Mateusz Kurciński

MODELOWANIE ODDZIAŁYWAŃ BIAŁEK I PEPTYDÓW W ZREDUKOWANEJ PRZESTRZENI KONFORMACYJNEJ

Praca doktorska wykonana
w Pracowni Teorii Biopolimerów
Wydziału Chemii
Uniwersytetu Warszawskiego

Promotor pracy:
Prof. dr hab. Andrzej Koliński

Warszawa 2014

Serdecznie dziękuję
mojemu promotorowi
profesorowi Andrzejowi Kolińskiemu
za ogromne wsparcie i pomoc
w trakcie realizacji tej pracy.

Dziękuję koleżankom i kolegom
z Pracowni Teorii Biopolimerów,
za współpracę i miła atmosferę.

W szczególności dziękuję
Sebastianowi Kmiecikowi
za cenne wskazówki i pomoc.

Dziękuję rodzinie,
w szczególności rodzicom
za nieustające wsparcie.

Pracę dedykuję mojej
wspaniałej żonie i córkom.

Spis Treści

I Cel Pracy 9

II Podstawy teoretyczne 13

1 Metody doświadczalne badania struktur kompleksów białkowych 13

1.1 Krystalografia 13

1.2 Spektroskopia NMR 14

2 Teoretyczne modelowanie struktur białkowych 15

3 Dokowanie 19

3.1 Dokowanie sztywne 20

3.2 Dokowanie giętkie 23

III CABSDock – opracowanie automatycznej procedury dokowania białko-białko w zredukowanej przestrzeni konformacyjnej. 29

1 Ogólny schemat procedury dokowania 30

2 Dokowanie sztywne – FTDOCK 32

2.1 Reprezentacja molekuł 32

2.2 Funkcja oceny 32

2.3 Przeszukiwanie przestrzeni konformacyjnej 33

3 Selekcja modeli do dokowania giętkiego 34

3.1 Obliczanie energii oddziaływania E_{inter} – CABScore 35

3.2 Analiza skupień 37

3.3 Ranking modeli 37

4 Dokowanie giętkie za pomocą modelu CABS 38

4.1 Opis modelu CABS 38

5 Selekcja i wygładzanie końcowych modeli 44

5.1 Analiza skupień 44

5.2 Odbudowa modeli do pełnoatomowej reprezentacji 45

5.3 Końcowe wygładzanie modeli 45

IV Modelowanie kompleksów białkowych – uzyskane wyniki 47

1 Dokowanie białko-peptyd 47

1.1 Modelowanie trójwymiarowych struktur kompleksów receptorów jądrowych i peptydów imitujących czynniki transkrypcyjne 47

1.2 Wieloskalowe modelowanie trójwymiarowych struktur kompleksów białkowych 50

1.3 Badanie mechanizmu aktywacji receptora retinoidów α (RXRα) przez cząsteczki kwasu 9-cis retinowego i koaktywatora TRAP220. 52
1.4 Badanie mechanizmu jednoczesnego zwijania się białka nieustrukturyzowanego pKID w trakcie tworzenia kompleksu z domeną KIX białka CREB. 55
1.5 Modelowanie zjawiska mimikry molekularnej gangliozydu GD2 przez grupę peptydów w kompleksie z przeciwciałem 14G2a. 58
2 Dokowanie białko-białko 60
2.1 Teoretyczny model kompleksu białkowego ludzkiej Telomerazy. 60
V Podsumowanie i wnioski końcowe 65
VI Prace cytowane 69
VII Prace stanowiące podstawę rozprawy 81
Praca I (P.I) 81
Praca II (P.II) 87
Praca III (P.III) 97
Praca IV (P.IV) 105
Praca V (P.V) 125
Praca VI (P.VI) 139

I Cel Pracy

Białka występujące w znakomitej większości znanych[1] organizmów składają się jedynie z 20 rodzajów aminokwasów. Liczba ta wydaje się niewielka, zważywszy na mnogość form i funkcji tych biomolekuł. Gdyby jednak policzyć, ile jest teoretycznie możliwych 100-aminokwasowych polipeptydów ($20^{100} \approx 1.27 \cdot 10^{130}$), zrozumiałe staje się, dlaczego tak wiele zadań natura powierzyła właśnie białkom. W którąkolwiek stronę ewolucja popchnie organizm, pomiędzy gigantyczna liczbą kombinacji ułożenia aminokwasów prawie na pewno znajdzie się taka, która po utworzeniu IV-rzędowej struktury będzie mogła realizować nowe zadania.

Białka zajmują centralną pozycję w funkcjonowaniu żywych organizmów i są obecne w niemal wszystkich procesach zachodzących zarówno wewnątrz komórek jak i poza nimi. Spośród ogromnej liczby funkcji, które realizują do najważniejszych należą:

- regulacja reakcji biochemicznych (enzymy – np. dehydrogenaza alkoholowa),
- kontrola ekspresji genów (aktywatory/represory – receptor witaminy D),
- przekazywanie sygnałów do wnętrza komórek (GPCRs – rodopsyna),
- przekazywanie sygnałów pomiędzy organami (hormony – insulina),
- rozpoznawanie i unieszkodliwianie drobnoustrojów (immunoglobuliny),
- transport i magazynowanie składników odżywczych (hemoglobina, mioglobina),
- funkcje strukturalne (kolagen, keratyna).

Białka mogą wypełniać tak wiele różnych zadań w organizmach dzięki swej unikalnej zdolności do przyłączania w selektywny sposób innych biomolekuł – cukrów, lipidów, kwasów nukleinowych czy innych białek. Najczęściej przyłączanie jakiegoś liganda zmienia przestrzenną strukturę proteiny, przez co zaczyna ona funkcjonować w odmienny sposób. Przykładowo, polimeraza RNA w formie niezwiązanej nie reaguje z otaczającymi ją nukleotydami. Gdy tylko jednak przyłączy

[1] w białkach niektórych rodzajów bakterii i archeowców występują inne niż 20 standardowe aminokwasy, np. pyrrolizyna.

się do specyficznego regionu w nici DNA, zmienia się jej struktura przestrzenna i rozpoczyna produkcja mRNA - pierwszy krok w procesie ekspresji genu.

Zrozumienie mechanizmów działania białek jest obecnie jednym z najważniejszych problemów stojących przed naukami biologicznymi. Z racji tak wielu funkcji, które pełnią białka w organizmach, stanowią one idealny cel dla działania leków, co stawia je również w centrum zainteresowania przemysłu farmaceutycznego. Obecnie produkowane leki wykorzystują jedynie kilkanaście receptorów jako cele swojego działania, a przecież ludzki organizm zawiera około 25 tys. genów, z których większość koduje jakieś białko. Zrozumienie ich funkcjonowania pozwoli w przyszłości na produkowanie wysoce selektywnych i skutecznych farmaceutyków.

Ponieważ działanie białek opiera się na tworzeniu kompleksów z innymi molekułami, poznanie trójwymiarowych struktur takich kompleksów stanowi niezbędny etap na drodze do zrozumienia mechanizmów ich działania. W ostatnich dwóch dekadach nastąpił znaczący rozwój eksperymentalnych metod rozwiązywania trójwymiarowych struktur białek. Jednak po latach stałego wzrostu liczby deponowanych w bazie PDB[2] struktur, nastąpiło spowolnienie. Obecnie do bazy PDB trafia około 7 tys. struktur rocznie, z czego wiele to struktury podobne do wcześniej już zdeponowanych. Przy utrzymaniu obecnego poziomu zaangażowania zespołów badawczych i nakładów finansowych, liczba ta nie będzie się znacząco zmieniać w przyszłości. Jednocześnie liczba znanych sekwencji białkowych wynosi już blisko 100 mln i dzięki szybkim, tanim i automatycznym metodom sekwencjonowania całych genomów rośnie z roku na rok coraz szybciej. Dodatkowo w ostatnich latach wysiłki genetyków są ukierunkowane na tworzenie map wzajemnych oddziaływań białek, co dodaje dodatkowy wymiar do skomplikowanego już wcześniej zagadnienia. Wydaje się, że w dającej się przewidzieć przyszłości, nie będzie możliwe rozwiązanie problemu poznania wzajemnych oddziaływań genów dla nawet najmniejszych genomów. Pozostawia to szerokie pole do działania dla metod teoretycznych.

Pod pojęciem dokowania białek kryje się problem przewidywania struktury przestrzennej kompleksu, przy danych strukturach jego izolowanych komponentów. Jak wszystkie interesujące problemy naukowe, ten również łatwo sformułować, lecz bardzo trudno rozwiązać. Już w 1978 roku Wodak i Janin (Wodak & Janin 1978)

[2] Protein Data Bank – ogólnoświatowa baza danych rozwiązanych struktur białkowych.

opisali procedurę automatycznego algorytmu dokowania inhibitora do bydlęcej trypsyny trzustkowej. Przez kolejnych 30 lat dokowanie urosło do ważnej dyscypliny naukowej łączącej elementy biologii, chemii, fizyki, matematyki i technik informatycznych we wspólnym wysiłku poznania mechanizmów funkcjonowania tak złożonych systemów biologicznych jak kompleksy białkowe. Rozwój tej dziedziny jest stale dokumentowany w projekcie CAPRI[3] i mimo znacznego postępu, jaki dokonał się zwłaszcza w ostatnich latach, problem dokowania daleki jest wciąż od zadowalającego rozwiązania. Główną przeszkodą wydaje się uwzględnienie konformacyjnej giętkości molekuł w trakcie formowania kompleksów. Znakomita większość dostępnych już algorytmów traktuje molekuły, jako bryły sztywne, a ich dopasowanie dokonuje się na podstawie komplementarności kształtów powierzchni molekularnych. W rzeczywistości białka potrafią znacznie zmienić swoja strukturę na etapie tworzenia kompleksów, co w oczywisty sposób nie może być odzwierciedlone w modelu przy założeniu sztywności molekuł.

Celem tej pracy było opracowanie metody teoretycznego przewidywania struktur kompleksów białkowych za pomocą modelu CABS, opracowanego przez Kolińskiego w 2004 roku (Kolinski 2004). CABS w pierwotnej postaci powstał jako narzędzie służące do przewidywania struktur pojedynczych białek. Na potrzeby obecnej pracy został zaadaptowany do symulowania dynamiki wielu łańcuchów polipeptydowych. Głównym założeniem opracowanej metody było zachowanie całkowitej giętkości symulowanych molekuł tak, aby możliwe było symulowanie procesów dokowania, w których dochodzi do znacznych zmian konformacji. Jednocześnie pragnieniem autora było całkowite zautomatyzowanie procesu dokowania, przez co opisany tutaj algorytm jest w rzeczywistości wielostopniową procedurą, składającą się z wielu niezależnych programów, które połączone zostały w automatyczne narzędzie realizujące dokowanie białko-białko.

[3] **C**ritical **A**ssessment of **PR**otein **I**nteractions - światowy konkurs przewidywania struktur kompleksów białkowych, w którym grupy teoretyków próbują przewidzieć ostatnio rozwiązane, lecz jeszcze nieujawnione struktury.

II Podstawy teoretyczne

1 Metody doświadczalne badania struktur kompleksów białkowych

1.1 Krystalografia

W 1958 roku opublikowano strukturę mioglobiny kaszalota rozwiązaną jako pierwszą za pomocą krystalografii rentgenowskiej (Kendrew et al. 1958). Obecnie około 90% struktur obecnych w bazie PDB zostało rozwiązanych tym sposobem. W metodzie tej kryształ białka jest oświetlany wiązką promieniowania X, która ulega dyfrakcji na atomach molekuły w specyficznych kierunkach (Brink et al. 1954; McPherson 1985). Analiza kątów i intensywności zarejestrowanych wiązek promieniowania pozwala na stworzenie mapy gęstości elektronowej, która za pomocą specjalistycznego oprogramowania tłumaczona jest na trójwymiarową strukturę cząsteczki (Otwinowski & Minor 1997). Główną wadą tej metody jest konieczność uzyskania odpowiedniego kryształu białka – wystarczająco dużego i wolnego od defektów (Stevens 2000). W ostatnich latach zaczęto stosować wyspecjalizowane roboty, które produkują dużą liczbę kryształów tego samego białka, zwiększając szansę na pomyślną krystalizację (Cox & Weber 1987). O ile dla białek globularnych uzyskiwanie kryształu stało się standardową procedurą (Usón & Sheldrick 1999), to krystalizacja białek membranowych pozostaje wciąż wymagającym i często zakończonym porażką procesem (Carpenter et al. 2008). Dodatkowym problemem, z którym spotykają się krystalografowie jest jakość map gęstości elektronowej. Dyfrakcja promieniowania na mobilnych fragmentach białka zachodzi w odmienny sposób w różnych komórkach elementarnych, co przekłada się na rozmazany obraz dyfraktogramu i niemożność jednoznacznego dopasowania struktury do mapy gęstości elektronowej (Jones et al. 1991). Zwiększenie rozdzielczości można osiągnąć stosując wiązki promieniowania o wysokiej energii, uzyskiwane jedynie w synchrotronach (Helliwell 1984). Konieczność zastosowania wysoce specjalistycznego sprzętu i wysokie ryzyko porażki na wielu etapach procesu sprawia, że uzyskiwanie struktury białek za pomocą krystalografii jest technologią bardzo kosztowną.

1.2 Spektroskopia NMR

Drugą metodą rozwiązywania struktur białkowych jest spektroskopia NMR. Około 10% struktur z bazy PDB zostało rozwiązanych tą metodą. Pionierami zastosowania NMR do rozwiązywania struktur białek byli Kurt Wütrich i Richard Ernst (Wuthrich 1969; Kumar et al. 1981). W metodzie tej wykorzystuje się magnetyczne własności jąder pewnych szczególnych atomów (głównie ^{1}H and ^{13}C). Umieszczone w polu magnetycznym jądro obdarzone spinem, po przyjęciu kwantu promieniowania o określonej częstości wpada w rezonans, co może być zarejestrowane przez detektor (Bloch 1946; Bloembergen et al. 1947; Bloembergen et al. 1948). Intensywność sygnału jest proporcjonalna do liczby równoważnych chemicznie jąder, a energia rezonansu zależy od otoczenia, w jakim znajduje się dane jądro. Teoretycznie każde jądro posiada inne otoczenie, przez co charakteryzuje się unikalną częstością rezonansu. W praktyce sygnały od wielu jąder nakładają się na siebie, przez co analiza widm NMR jest zadaniem nietrywialnym i wymaga zaangażowania zaawansowanych narzędzi i wysokiej klasy specjalistów (Bax & Grzesiek 1993; Herrmann et al. 2002). Wraz ze wzrostem rozmiaru badanej molekuły nakładanie się sygnałów stanowi coraz większy problem, co ogranicza zastosowanie tej metody jedynie do niewielkich białek. Tak jak w przypadku krystalografii, spektroskopia NMR jest bardzo kosztowną metodą uzyskiwania struktur białek, zarówno ze względu na specjalistyczny sprzęt jak i konieczność zaangażowania wysoce wykwalifikowanej obsługi (Qureshi & Goto 2012; Chou & Sounier 2013).

Krystalografia rentgenowska i spektroskopia NMR są obecnie jedynymi metodami pozwalającymi na uzyskanie struktur białek o rozdzielczości pozwalającej na ich praktyczne zastosowanie. Na dzień dzisiejszy za ich pomocą udało się uzyskać blisko 100 tys. struktur podczas gdy sekwencji białkowych znanych jest już około 100 mln. W 2000 roku z budżetem przekraczającym 700 mln dolarów wystartował szeroko zakrojony projekt o nazwie „Protein Structure Initiative" (Stevens et al. 2001). Naukowcy z kilkudziesięciu laboratoriów badawczych postawili sobie za cel redukcję nakładów finansowych i czasowych potrzebnych do uzyskania struktur białek. Przez prawie 10 lat udało im się rozwiązać prawie 3500 nowych struktur (Berman et al. 2009; Gifford et al. 2012),w tym około 2000 unikalnych. Przy stałym zaangażowaniu nawet tych rekordowych jak dotychczas środków finansowych i personalnych poznanie struktur wszystkich 100 mln obecnie znanych sekwencji zajęłoby tysiące lat i kosztowało wiele bilionów dolarów. Nawet tak naiwne przybliżenie oddaje skalę problemu, przed którym stoi dziś genomika strukturalna.

2 Teoretyczne modelowanie struktur białkowych

Ogromna przepaść pomiędzy liczbą znanych sekwencji i struktur białkowych wymusza rozwój szybkich i tanich metod przewidywania tych ostatnich. Pierwsze modele opisujące dynamikę białek powstały ponad 30 lat temu i zorientowane były przede wszystkim na modelowanie kompletnego procesu zwijania (Levitt & Warshel 1975; Warshel & Levitt 1976). W 2014 roku autorzy tych modeli otrzymali nagrodę Nobla. Poznanie ścieżki zwijania pozostaje głównym celem modelowania białek, lecz pomimo ogromnego postępu wciąż jest to zadanie nietrywialne. Natomiast przewidywanie jedynie końcowej struktury białka stanowi dla niego atrakcyjną alternatywę z dwóch powodów. Po pierwsze, dla celów praktycznych (np. projektowania leków) często wystarcza jedynie znajomość stabilnej struktury białka. Ponadto skoncentrowanie się jedynie na końcowej strukturze pozwala w pełni na wykorzystanie ogromnej ilości informacji pochodzących z rozwiązanych już struktur.

Podstawą współczesnych metod modelowania struktur białkowych jest obserwacja, że ewolucyjnie powiązane białka posiadają podobne struktury (Chothia & Lesk 1986). Dzięki temu faktowi możliwe jest modelowanie porównawcze – modelowanie z wykorzystaniem znanych struktur bądź ich fragmentów jako szablonów do przewidywania nowych, o ile ich sekwencje wykazują istotne statystycznie podobieństwo (Sánchez & Šali 1997; Dunbrack 2006). W zależności od stopnia podobieństwa sekwencji rozwiązywanego białka i sekwencji szablonu problem modelowania struktury można zakwalifikować do jednej z trzech kategorii:

- **modelowanie porównawcze** – Podobieństwo sekwencyjne powyżej 35% wskazuje na pokrewieństwo ewolucyjne białek w 90% przypadków (Sander & Schneider 1991). Modelowanie z użyciem szablonu będącego bliskim homologiem pozwala na uzyskanie pełnoatomowego modelu wysokiej jakości, porównywalnego do modeli eksperymentalnych. Obecnie modelowanie homologiczne wykonywane jest w pełni automatycznie.
- **przewlekanie**[4] – W przypadku braku szablonów będących bliskimi homologami modelowanego białka, jako szablony wykorzystuje się fragmenty struktur, które wykazują nawet niewielkie podobieństwo do modelowanego białka. W tym celu „przewleka" się modelowaną sekwencję przez znane struktury i ocenia za pomocą funkcji przybliżających energię swobodną (Bowie et al. 1991). Jakość modeli

[4] (ang.) threading

uzyskanych tą metodą istotnie zależy od identyfikacji właściwych szablonów oraz od dokładnego uliniowienia modelowanej sekwencji do sekwencji szablonów. Wyniki kolejnych edycji eksperymentu CASP pokazują, że w tej kategorii doświadczony ekspert wciąż posiada pewną przewagę nad automatycznymi procedurami (Kinch et al. 2011).

- **modelowanie *de novo*** – Modelowanie bez użycia szablonu na dzień dzisiejszy pozwala na uzyskanie struktur jedynie niewielkich białek (Kinch et al. 2011). Ograniczenia tej metody wynikają z konieczności przeszukania ogromnej przestrzeni konformacyjnej białka. Dokładne metody bazujące na pełnoatomowej dynamice molekularnej są zbyt kosztowne obliczeniowo, aby mogły być stosowane dla całych genomów. Najskuteczniejsze w tej kategorii są obecnie modele zredukowane o wysokiej rozdzielczości (Rohl et al. 2004; Koliński & Bujnicki 2005; Zhang 2008).

Białka w stanie natywnym znajdują się w globalnym minimum energii swobodnej (Anfinsen 1973). Energia swobodna jest termodynamiczną funkcją stanu i na jej wartość nie wpływają stany pośrednie na ścieżce zwijania. Zatem niezależnie od faktycznej drogi jaką musi pokonać białko *in vivo* do stanu natywnego, teoretyczne przewidywanie struktury sprowadza się do konstrukcji funkcji jak najlepiej przybliżającej energię swobodną i efektywnej metody poszukiwania globalnego minimum tej funkcji.

Przez ostatnice 30 lat opracowano wiele metod służących do rozwiązania tak postawionego problemu. Pomimo ogromnego postępu jaki się w tym czasie dokonał, wciąż głównym wyzwaniem w tej dziedzinie pozostaje znalezienie kompromisu pomiędzy dokładnością z jaką model reprezentuje przestrzeń konformacyjną układu a efektywnością jej przeszukiwania. Podejście kwantowo-mechaniczne, które najdokładniej oddaje krajobraz energetyczny, nie pozwala na symulowanie układów wielkości białka. Warshel i Levitt zaproponowali połączenie mechaniki kwantowej i mechaniki klasycznej w metodzie QM/MM[5] (Warshel & Levitt 1976), w której modelowany układ zostaje podzielony na dwie części. Część odpowiedzialna za kluczowe oddziaływania (np. z ligandem) traktowana jest w sposób kwantowo-mechaniczny, natomiast pozostałe fragmenty białka wraz z molekułami rozpuszczalnika traktowane są w sposób klasyczny.

Kolejną metodą modelowania układów białkowych jest Dynamika Molekularna. Metoda ta została opracowana na przełomie lat 50-tych i 60-tych XX wieku (Alder &

[5] QM/MM – Quantum Mechanics/Molecular Mechanics

Wainwright 1959; Rahman 1964). Polega na rozwiązywaniu równań ruchu Newtona dla atomów modelowanego układu i ich ewolucji w czasie. Dynamika Molekularna została po raz pierwszy zastosowana do badania białek przez Michaela Levitt'a (Levitt et al. 1985) i do dzisiaj pozostaje „złotym standardem" modelowania dynamiki i termodynamiki białek, gdyż pozwala na odtworzenie detali struktury w stopniu wystarczającym do większości praktycznych zastosowań. Minusem Dynamiki Molekularnej jest jej koszt obliczeniowy, co przekłada się na niewielki zakres dostępnej skali czasowej modelowanych procesów. Symulacja Dynamiki Molekularnej typowej wielkości białka przy zastosowaniu obliczeń równoległych na klastrze obliczeniowym osiąga obecnie prędkości rzędu 100 ns/dzień (Saibil et al. 2009), natomiast czas potrzebny *in vivo* na zwinięcie się białek jest rzędu milisekund (Lindorff-Larsen et al. 2011), a na utworzenie się kompleksów białkowych często potrzeba minut lub nawet godzin (Bae et al. 2013).

W odpowiedzi na ograniczony zakres zastosowania Dynamiki Molekularnej w badaniu białek rozwinęła się klasa metod wykorzystujących dynamikę Monte Carlo. Metodę tę opracowali Ulam i Metropolis (Metropolis & Ulam 1949) w trakcie prac nad projektem „Manhattan". Autorzy zauważyli, że do rozważania wielu problemów termodynamiki statystycznej korzystniej jest modelować ewolucję badanego układu w sposób stochastyczny, a nie deterministyczny. Wraz z rozwojem maszyn obliczeniowych metoda Monte Carlo znalazła zastosowanie w modelowaniu dynamiki układów molekularnych, w tym białek (Tanaka & Scheraga 1975; Kolinski et al. 1986; Kolinski & Skolnick 1994a; Kolinski & Skolnick 1994b; Kolinski 2004; Rohl et al. 2004). Przewagą metody Monte Carlo nad Dynamiką Molekularną jest brak konieczności całkowania równań ruchu. W dalszym kroku przekłada się to na możliwość wykorzystania w dynamice Monte Carlo niecałkowalnych funkcji energii – w szczególności potencjałów statystycznych, co w jeszcze większym stopniu obniża koszt obliczeniowy. W efekcie symulacje dynamiki Monte Carlo, w stosunku do Dynamiki Molekularnej, pozwalają na modelowanie zarówno większych układów oraz procesów o dłuższych skalach czasowych, przy porównywalnym koszcie obliczeniowym.

Kolejnym zabiegiem mającym na celu efektywniejsze przeszukiwanie przestrzeni konformacyjnej układu jest zredukowanie jej wymiarów. W Dynamice Molekularnej układ posiada 3N-6 stopni swobody, gdzie N jest całkowitą liczbą atomów układu, włączając atomy wchodzące w skład cząsteczek rozpuszczalnika. Dla typowej wielkości białka, w symulacji w pudle z jawnym rozpuszczalnikiem, N jest liczbą rzędu 10^4, z czego ok 90% stanowią atomy rozpuszczalnika. Liczbę tę można istotnie zredukować stosując model niejawnie

uwzględniający wpływ rozpuszczalnika, w miejsce rzeczywistych atomów. Typowo stosuje się modele oparte na równaniu Poissona-Boltzmana (Fogolari et al. 2002; Grochowski & Trylska 2008), które traktują molekułę białka i rozpuszczalnik jako dwa obszary o różnych stałych dielektrycznych. W miejsce oddziaływań białka z poszczególnymi cząsteczkami rozpuszczalnika, uwzględnia się jedynie oddziaływanie z potencjałem średniej siły pochodzącej od otoczenia.

Innym sposobem na zmniejszenie liczby wymiarów przestrzeni konformacyjnej układu jest zredukowanie liczby traktowanych w sposób jawny atomów w samej molekule białka. Łańcuch polipeptydowy jest relatywnie sztywny – wiele stopni swobody jest w dużym stopniu zależnych od siebie i w dosyć naturalny sposób możliwe jest w modelu reprezentowanie całej grupy atomów przez jeden z nich bądź zastąpienie przez hipotetyczny pseudoatom. Model Wilsona i Doniacha wykorzystywał pojedynczy atom węgla Cα do reprezentacji całego aminokwasu w symulacji zwijania Krambiny (Wilson & Doniach 1989). Podobne podejście zaprezentowano w pracy (Trylska et al. 2005), w której symulowano ruchy całego rybosomu. Odmienne podejście zaprezentowano w modelu SICHO (Kolinski & Skolnick 1998), gdzie również cały aminokwas reprezentowany był przez jeden pseudoatom, lecz zlokalizowany w geometrycznym środku ciężkości łańcucha bocznego. Model UNRES (Liwo et al. 1998) wykorzystuje po dwa centra oddziaływań na jedną resztę, przy czym centrami nie są punktowe atomy, lecz elipsoidy, których rozmiary i orientacja przestrzenna odpowiadają grupom atomów, które reprezentują. W modelu ROSETTA (Rohl et al. 2004) łańcuch boczny pojedynczego aminokwasu jest zredukowany do centroidu, a jedynie na potrzeby pewnych członów funkcji energii wykorzystuje się pełnoatomową reprezentację. Natomiast w modelu CABS (Kolinski 2004) na każdą resztę aminokwasową przypada do czterech centrów oddziaływań: atom Cα, atom Cβ (oprócz glicyny), pseudoatom łańcucha bocznego (oprócz glicyny i alaniny) oraz pseudoatom wiązania peptydowego.

Dzięki zastosowaniu modeli zredukowanych do modelowania homologicznego stało się ono obecnie standardową procedurą badawczą, wykonywaną w pełni automatycznie. Identyfikacja i uliniowienie szablonu jest łatwe i szeroko dostępne dzięki publicznym metaserwerom (Bujnicki et al. 2001; Kosinski et al. 2003). Szacuje się, że dla około 60% znanych sekwencji białkowych możliwe jest uzyskanie struktur o rozdzielczości bliskiej krystalograficznej (Pieper et al. 2004). Rozwiązane teoretycznie struktury znajdują już dziś szereg zastosowań (Saibil & Zhang 2009), jednak wciąż ustępują jakością tym uzyskanym eksperymentalnie. Dlatego też obecnie wiele grup badawczych koncentruje swoje wysiłki na opracowaniu metod udokładniania modeli teoretycznych (Teichmann & Zhang 2008). Mimo

że problem zwijania się białek do struktury przestrzennej daleki jest jeszcze od pełnego rozwiązania, to obecny stan wiedzy na ten temat pozwala na przejście do kolejnego etapu rozwikływania kodu genetycznego – badania oddziaływań białek.

3 Dokowanie

Dokowanie polega na teoretycznym przewidzeniu przestrzennej struktury kompleksu białkowego, znając jedynie struktury jego izolowanych komponentów. Na przestrzeni 30 lat, odkąd po raz pierwszy podjęto próby modelowania oddziaływań międzybiałkowych, powstało wiele metod realizujących to zadanie na różne sposoby. Pierwszego opisu zjawiska łączenia się białek z biomolekułami dostarczał historyczny już model „zamka-klucza" (Fischer 1894). Model ten zakładał, że cząsteczka liganda w konkretnej konformacji „pasuje" do kieszeni na powierzchni molekuły receptora, tak jak klucz pasuje do zamka. Kolejny model nazwany „indukowanym dopasowaniem" (Koshland 1958) zakładał, że molekuła liganda dopasowuje swoją konformację do kształtu kieszeni na powierzchni receptora. Oba te poglądy zostały zweryfikowane, gdy w bazie PDB zaczęły pojawiać się pierwsze struktury kompleksów, dla których znane były również struktury wykrystalizowanych osobno komponentów, często różniące się znacznie w swojej formie niezwiązanej od formy skompleksowanej. Wiedząc również, że białka łączą się także między sobą, jasnym stało się, że zmiana konformacji nie może zachodzić tylko w jednej z łączących się molekuł. Aktualny obecnie opis zjawiska tworzenia kompleksów białkowych, nazwany „konformacyjną selekcją" (Boehr et al. 2009), zakłada, że zarówno receptor jak i ligand występują jednocześnie w wielu konformacjach, z których najkorzystniejsze energetycznie prowadzą do wytworzenia kompleksu, po którym następuje redystrybucja populacji pomiędzy konformacjami. W efekcie obserwuje się wzajemne dopasowanie konformacji oddziałujących cząsteczek, często wykraczające daleko poza miejsce kontaktu na powierzchni oddziaływania. Przykładowo: dla receptora retinoidów α wartość RMSD[6] pomiędzy strukturą w formie

[6] RMSD - **R**oot **M**ean **S**quare **D**eviation to wielkość statystyczna, często używana jako miara podobieństwa strukturalnego molekuł. RMSD definiuje się jako pierwiastek kwadratowy z uśrednionych kwadratów odległości pomiędzy odpowiadającymi sobie atomami w dwóch strukturach.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|X_i - Y_i\|^2} \qquad [2.1]$$

Ponieważ w takiej definicji wartość RMSD zależy od względnego ustawienia molekuł, obliczanie RMSD poprzedza się procedurą ustawienia cząsteczek względem siebie tak, aby RMSD było minimalne.

niezwiązanej (kod PDB: 1LBD), a formą związaną z kwasem retinowym (kod PDB: 2LBD) wynosi prawie 8Å.

Z punktu widzenia mechaniki kwantowej oddziaływania, dzięki którym białko łączy się z innymi biomolekułami (małą cząsteczką organiczną, kwasem nukleinowym, cukrem, lipidem lub innym białkiem) są tej samej natury. Jednakże w miarę rozwoju metod dokowania problem został rozbity na kilka klas, w których podejście do modelowania jest zupełnie inne. W przypadku, gdy receptorem jest molekuła białka a ligandem mała cząsteczka organiczna, mamy do czynienia z tak zwanym dokowaniem białko–lek. Nazwa ta sugeruje od razu najważniejsze zastosowanie dla takich metod. Gdy zarówno receptor jak i ligand są białkami, mówi się o dokowaniu białko-białko - obecnie najszybciej rozwijającej się klasie metod. Dokowanie białek do kwasów nukleinowych, błon lipidowych i innych bioagregatów jest na razie poza zasięgiem współczesnych metod obliczeniowych i trudno mówić o jakichś standardowych podejściach do tego problemu. W tej pracy skoncentrowano się na przedstawieniu aktualnego stanu metod dokowania białko-białko oraz pewnym szczególnym przypadku dokowania białko-lek, gdy ligandem jest krótki peptyd.

3.1 Dokowanie sztywne

W 1969 roku Levinthal opublikował pracę, w której przedstawił swój słynny paradoks (Levinthal 1969), który głosił, że gdyby białka miały się zwijać do swoich form natywnych poprzez sekwencyjne próbkowanie wszystkich możliwych konformacji, proces taki trwałby dłużej niż wiek wszechświata. Powodem tego jest astronomiczna liczba stopni swobody typowego łańcucha białkowego. Podobne rozumowanie można zastosować do problemu tworzenia przez białka kompleksów. Nawet gdyby komputery potrafiły próbkować kolejne konformacje równie szybko, jak zachodzi to *in vivo* to i tak nie byłoby możliwe sprawdzenie wszystkich możliwych konformacji układu o 3N-3 stopniach swobody, a tyle ich właśnie posiada układ N atomów traktowanych w całkowicie giętki sposób. Przy założeniu całkowitej sztywności zarówno cząsteczek receptora jak i liganda przestrzeń konformacyjna układu posiada jedynie sześć wymiarów. Dokowanie sprowadza się wtedy do znalezienia trzech składowych wektora przesunięcia i trzech kątów obrotu jednej z molekuł względem drugiej.

Tak jak prawie wszystkie procesy biochemiczne również łączenie się białek w kompleksy jest najczęściej dyktowane obniżeniem całkowitej energii swobodnej układu. Zagadnienie dokowania jest zatem szczególnym przypadkiem szerokiej klasy problemów

poszukiwania minimum globalnego na hiperpowierzchni pewnej funkcji oceny, która tę energię przybliża. Metody realizujące przeszukiwanie przestrzeni konformacyjnej *in sillico* można podzielić na trzy klasy:

- ukierunkowane, gdzie ewolucja układu zachodzi w sposób deterministyczny (np. metoda największego spadku);
- stochastyczne, gdzie układ poddawany jest losowym zaburzeniom, z których wybierane są te, które prowadzą układ w kierunku minimum (np. metoda Monte Carlo);
- wyczerpujące, gdzie sprawdzane są wszystkie możliwe konformacje układu i wybierana jest najlepsza.

Dla układów o wielu stopniach swobody, metody z ostatniej z wyżej wymienionych klas są zbyt kosztowne obliczeniowo, aby znaleźć praktyczne zastosowanie. Natomiast w przypadku jedynie sześciowymiarowej przestrzeni konformacyjnej, wykorzystanie wyczerpującego przeszukiwania jest zadaniem jak najbardziej w zasięgu współczesnych komputerów. Istnieje szereg programów realizujących dokowanie sztywne w taki właśnie sposób. Poniżej przedstawiono ich najważniejsze cechy.

3.1.1 Reprezentacja molekuł

Przy założeniu ciągłej reprezentacji przestrzeni konformacyjnej liczba możliwych konformacji jest nieskończona. Dlatego konieczne jest wprowadzenie przybliżonej reprezentacji – modeli gruboziarnistych. W większości metod dokowania sztywnego atomy są rzutowane na siatkę, której charakterystyka (typ siatki, rozdzielczość) jest zazwyczaj parametrem wybieranym przez użytkownika. W klasycznym podejściu do dokowania sztywnego, gdzie dopasowanie wykorzystuje komplementarność kształtów cząsteczek, molekuły receptora i liganda rzutowane są na osobne siatki w binarnej procedurze: każdemu węzłowi siatki przypisywana jest wartość 1, jeżeli przynajmniej jeden atom znajduje się w odległości nie większej niż r (gdzie r jest rzędu promienia van der Waalsa atomu) od węzła. Pozostałym węzłom przypisywana jest wartość 0. Tak zdefiniowane siatki są następnie nakładane na siebie na wszystkie możliwe sposoby - sprawdzane są wszystkie translacje i rotacje pomiędzy nimi.

3.1.2 Przeszukiwanie przestrzeni konformacyjnej

Nawet dla sześciowymiarowej przestrzeni konformacyjnej układu, przy rozsądnych wartościach stałej siatkowej [7], liczba konformacji do sprawdzenia jest rzędu 10^9. W 1992 roku Katchalski-Katzir zaproponował bardzo efektywne i jednocześnie eleganckie rozwiązanie problemu wyczerpującego przeszukiwania za pomocą techniki FFT[8] (Katchalski-Katzir et al. 1992). Po zrzutowaniu molekuł na siatkę dla każdej możliwej rotacji liganda (mniejszej cząsteczki) względem receptora (większej cząsteczki) obliczana jest funkcja korelacyjna za pomocą FFT. W ten sposób wszystkie możliwe translacje pomiędzy siatkami receptora i liganda zostają sprawdzone w czasie rzędu $o(N^3 ln(N^3))$, zamiast $o(N^6)$ gdyby FFT nie było stosowane. Znakomita większość metod sztywnego dokowania wykorzystuje powyższy schemat lub jego różne warianty.

3.1.3 Funkcja oceny

Pierwsze programy stworzone do sztywnego dokowania oceniały wiarygodność wygenerowanych kompleksów jedynie na podstawie dopasowania kształtów powierzchni molekularnych. Z biegiem czasu funkcje oceniające były wzbogacane o kolejne człony odpowiedzialne za komplementarne dopasowanie tak zwanych punktów krytycznych na powierzchniach cząsteczek - na przykład atomów obdarzonych cząstkowym ładunkiem dodatnim do atomów z ładunkiem ujemnym, donorów protonu do akceptorów itp. W chwili obecnej dominuje trend rozbicia oceny generowanych kompleksów na dwa etapy – ocen cząstkowych i rankingu (Warren et al. 2006; Krovat et al. 2005; Moitessier et al. 2008). W trakcie wyczerpującego skanowania wszystkich możliwych orientacji molekuł sprawdzana jest jedynie komplementarność kształtów z ewentualnym uwzględnieniem prostych oddziaływań (np. elektrostatyki). Na tym etapie odrzucane są struktury, w których komponenty kompleksu nakładają się na siebie lub wcale nie oddziałują. Pozostałe struktury szeregowane są względem stopnia oddziaływania, a następnie jedynie pewna ich liczba (parametr metody, zazwyczaj rzędu 10^4) jest zachowywana do dalszych etapów. Następnie

[7] Dla siatki prostokątnej najczęściej stosuje się stałą siatkową równą 0.5-1.0 Å, co stanowi kompromis między dokładnością a kosztem obliczeń. Dla wartości większych od 1 Å rzutowanie na siatkę wprowadza istotny błąd - typowa długość wiązania chemicznego to ok. 1.5 Å. W przypadku stałej siatkowej mniejszej od 0.5 Å koszt obliczeń istotnie wzrasta, a zmniejszenie błędu rzutowania jest tylko pozorne, gdyż w takim wypadku jest on często mniejszy niż rozdzielczość eksperymentalnie uzyskanych struktur.

[8] **F**ast **F**ourier **T**ransformation - Szybka Transformacja Fouriera

struktury są ponownie oceniane, tym razem za pomocą bardziej zaawansowanej funkcji oceny, uwzględniającej wiele typów oddziaływań pomiędzy atomami receptora i liganda.

Podsumowując, mimo swoich ograniczeń metody sztywnego dokowania stanowią ważną klasę technik teoretycznego przewidywania struktur kompleksów białkowych (Krovat et al. 2005; Warren et al. 2006). Do ich głównych zalet należą uniwersalność (uwzględniając tylko kształt cząsteczek nadają się do dokowania zarówno białek jak i innych biomolekuł) i niski koszt obliczeniowy (czas potrzebny na wykonanie dokowania metodą FFT dla średniej wielkości kompleksu wynosi kilka godzin na pojedynczym procesorze). Z racji braku lub bardzo ograniczonej giętkości cząsteczek, nie jest możliwe jednak przewidywanie tą techniką struktur kompleksów, w których następuje istotna zmiana konformacji w trakcie łączenia się komponentów. Dodatkowo istotnym ograniczeniem metod sztywnego dokowania jest konieczność ponownego uszeregowania uzyskanych struktur zgodnie z bardziej zaawansowanymi funkcjami oceny, których koszt obliczeniowy może wielokrotnie przekraczać koszt samego dokowania. Niemniej jednak metody sztywnego dokowania są powszechnie wykorzystywane na początkowym etapie modelowania, jako źródło wstępnych struktur o niskiej rozdzielczości (Halperin et al. 2002).

3.2 Dokowanie giętkie

Pierwsze modele opisujące proces łączenia się białek w kompleksy zakładały, że już w formie niezwiązanej molekuła receptora jest gotowa na przyłączenie cząsteczki liganda (Fischer 1894; Koshland 1958). Z tego założenia wynikało, że struktura receptora może tylko nieznacznie zmienić się w trakcie tworzenia kompleksu, gdyż w przeciwnym wypadku cząsteczka liganda nie byłaby w stanie rozpoznać receptora z którym miałaby się połączyć. Obecnie wiadomo już, że model ten dostarczał niedokładnego opisu zjawiska łączenia się białek, gdyż znanych jest wiele przykładów kompleksów, których komponenty wykrystalizowano zarówno w formie swobodnej, jak i skompleksowanej, wykazujących istotne różnice strukturalne. Obowiązujący dzisiaj model zakłada dwustopniowy proces rozpoznania/łączenia się białek w kompleksy (Vajda et al. 2013). W pierwszym etapie molekuły białek zaczynają przyciągać się za pomocą tak zwanych reszt-kotwic [9], zlokalizowanych na powierzchni cząsteczek w pobliżu, choć wcale niekoniecznie na samej powierzchni kontaktu (Rajamani et al. 2004). Resztami-kotwicami są zazwyczaj naładowane

[9] (ang.) anchor-residues

elektrostatycznie aminokwasy: lizyna, arginina, kwas glutaminowy i kwas asparaginowy. Dzieje się tak, ponieważ jedynie oddziaływania kulombowskie rozciągają się na tyle daleko w przestrzeni, aby cząsteczki mogły się efektywnie rozpoznać i ustawić we właściwej orientacji. Na bliskich odległościach między atomami znaczenie przyciągania elektrostatycznego maleje w stosunku do oddziaływań hydrofobowych[10], wspomaganych dodatkowo przez tworzące się wiązania wodorowe. O ile na etapie rozpoznania, komplementarność kształtów powierzchni molekularnych nie stanowi istotnego czynnika, o tyle w momencie połączenia się cząsteczek bardzo ważne jest precyzyjne umiejscowienie oddziałujących ze sobą atomów (np. donorai akceptora protonu). Modelując strukturę kompleksu za pomocą dokowania sztywnego, ligand może "nie pasować" do receptora, gdy obie molekuły znajdują się w swoich konformacjach, w których mają się jedynie rozpoznać i przyciągnąć. Dopiero wzajemnie indukowana zmiana konformacji molekuł pozwala na utworzenie kompleksu, co w oczywisty sposób nie może być odtworzone za pomocą symulacji dokowania sztywnego.

Główną trudnością wynikającą z wprowadzenia giętkości do algorytmu dokowania jest ogromna liczba stopni swobody symulowanego układu, co przekłada się na dużo wyższy niż w przypadku dokowania sztywnego koszt obliczeniowy. W chwili obecnej nie istnieją narzędzia, które byłyby w stanie przeprowadzić dokowanie w całkowicie giętki sposób, przy zachowaniu pełno-atomowej reprezentacji molekuł.

Prawie 50 lat temu Moore opublikował artykuł (Moore 1965), w którym przedstawił swoje empiryczne prawo postulujące wykładniczy wzrost liczby tranzystorów w komercyjnie dostępnych procesorach. Mimo wielokrotnych zapowiedzi kresu jego obowiązywania, cały czas pozostaje ono w mocy[11]. Od wynalezienia układów scalonych wysiłki inżynierów odpowiedzialnych za ich projektowanie koncentrowały się na minimalizacji rozmiarów tranzystorów tak, aby zmieściło się ich jak najwięcej na jednostce obliczeniowej. Minimalizacja tranzystorów musi mieć jednak kiedyś kres, dyktowany efektami kwantowymi (tunelowanie) oraz koniecznością coraz wydajniejszego chłodzenia procesorów. Z drugiej strony, dominujący ostatnio trend konstrukcji jednostek obliczeniowych, polegający na wykorzystaniu wielu rdzeni obliczeniowych, pozwala przypuszczać, że prawo Moore'a jeszcze przez pewien czas pozostanie niezagrożone.

[10] Energia oddziaływania elektrostatycznego jest odwrotnie proporcjonalna do kwadratu odległości pomiędzy ładunkami, podczas gdy przyciąganie hydrofobowe maleje proporcjonalnie do r^{-6}.

[11] Pierwotnie Moore postulował podwojenie liczby tranzystorów w jednostce obliczeniowej w ciągu 18 miesięcy. Obecnie, po dopasowaniu trendu do danych z ostatnich 40 lat okazało się, że liczba tranzystorów w procesorze podwajała się średnio co dwa lata.

Można zakładać, że wraz ze wzrostem mocy obliczeniowej komputerów, zostaną opracowane metody giętkiego dokowania, o dużo mniejszym niż obecnie stopniu uproszczenia reprezentacji molekuł. Jednak wydaje się, że przy obecnej technologii konstrukcji procesorów, opartej na wykorzystaniu krzemu, metody takie jak Dynamika Molekularna nie znajdą użytecznego zastosowania w dokowaniu giętkim. Proces łączenia się białek w kompleksy trwa w komórce od milisekund do minut, czy nawet lat (formacja amyloidów), podczas gdy najbardziej wydajne algorytmy Dynamiki Molekularnej pozwalają obecnie na symulację układu wielkości typowego białka w skali, co najwyżej nanosekundowej. Aby zatem skutecznie symulować tworzenie kompleksów białkowych z zachowaniem giętkości cząsteczek konieczne jest stosowanie modeli uproszczonych. Poniżej przedstawiono różne sposoby uwzględnienia giętkości molekuł w symulacjach dynamiki białek.

3.2.1 "Miękkie" dokowanie

Opisany w sekcji 3.1 schemat postępowania charakteryzuje typowy algorytm sztywnego dokowania, gdzie na żadnym etapie nie następuje zmiana konformacji wewnętrznej molekuł, a jedynie ich wzajemnej orientacji. Nie mniej jednak, nawet w sztywnym dokowaniu istnieje możliwość uwzględnienia giętkości cząsteczek, limitowana do grup bocznych na powierzchni molekuł, poprzez zastosowanie tak zwanej "miękkiej" funkcji potencjału SPF[12]. Na etapie rzutowania struktur na siatki, zamiast przypisywania węzłom dwóch wartości: 1 - molekuła, 0 - otoczenie, stosuje się dodatkowo rozgraniczenie na wnętrze cząsteczki oraz jej powierzchnię. Grubość warstwy oznaczającej powierzchnię molekularną jest zazwyczaj parametrem metody. Dzięki takiemu zabiegowi możliwe jest wprowadzenie dodatkowego stanu, który można przypisać strukturze wygenerowanego kompleksu. Oprócz "dopasowania", "braku oddziaływań" i "nakładania" z tej ostatniej grupy można wyodrębnić stan "penetracji", w którym węzły zdefiniowane jako powierzchnia nakładają się częściowo. W klasycznej reprezentacji taki układ zostałby odrzucony, jako posiadający zawady steryczne, a dzięki zastosowaniu SPF może być on włączony do zestawu struktur, w których nastąpiło dopasowanie. Zaletą takiego postępowania jest możliwość wygenerowania dokładniejszych modeli, wadą natomiast to, że tak wygenerowane struktury mogą zostać odrzucone przez funkcję oceny z powodu możliwego nakładania się atomów,

[12]SPF - **S**oft **P**otential **F**unction

skutkującego wysoką energią w większości pól siłowych. Oczywiście podejście takie nie pozwala na odtworzenie zmian konformacyjnych łańcucha głównego białka. Mimo wszystko „miękkie" potencjały są wciąż szeroko stosowane (Jiang & Kim 1991; Claußen et al. 2001; Carlson 2002; Barril & Fradera 2006; Moitessier et al. 2008; Cozzini & Kellogg 2008).

3.2.2 Dokowanie do zbioru struktur

Giętkość dokowanych cząsteczek może być częściowo uwzględniona w symulacjach wykorzystujących narzędzia do dokowania sztywnego, poprzez zastosowanie następującej procedury. W pierwszej kolejności generowany jest zbiór struktur różniących się konformacjami. Różnice mogą występować zarówno na poziomie rotamerów grup bocznych łańcucha peptydowego, lokalnych modyfikacji łańcucha głównego, jak również rozciągać się globalnie na całą molekułę białka. W przypadku gdy ligandem jest mała cząsteczka organiczna, zbiór taki zawiera zazwyczaj kolekcję różnych jej konformacji. Sposób ten pozwala na selektywne uwzględnienie giętkości. Można badać giętkość jedynie liganda przy zachowaniu sztywności receptora, jak i odwrotnie oraz obu cząsteczek na raz. Źródłem struktur mogą być alternatywne modele pochodzące z eksperymentów NMR, różne struktury krystalograficzne tego samego białka, jak również modele teoretyczne – na przykład z obliczeń Dynamiki Molekularnej lub ENM[13] (Bahar et al. 2007; Bakan & Bahar 2009; Tobi & Bahar 2005; Boehr et al. 2009). W kolejnym kroku przeprowadza się sztywne dokowanie pomiędzy wszystkimi parami receptor-ligand z każdego ze zbiorów, a następnie uzyskane modele szereguje za pomocą funkcji oceny.

Podstawową zaletą takiego podejścia do giętkiego dokowania jest niewątpliwie prostota implementacji, a także łatwość stosowania. Szeroka grupa narzędzi stworzonych pierwotnie do dokowania sztywnego może być rozszerzona o taką definicję giętkości (Leach 1994; Alberts et al. 2005; Carlson 2002; Barril & Fradera 2006; Moitessier et al. 2008). Jak łatwo jednak zauważyć, głównym minusem opisywanych metod jest koszt obliczeniowy, lawinowo rosnący wraz z liczbą wygenerowanych wstępnie konformerów. Ponadto wymienione metody dostarczają jedynie statycznego opisu zjawiska tworzenia kompleksów białkowych, niewystarczającego do badania ich mechanizmów.

[13] ENM - Elastic Network Model

3.2.3 Dynamika Molekularna

Dynamika Molekularna jest zbyt kosztowna obliczeniowo, aby możne ją było w praktyce stosować do dokowania *ab initio*. Czas potrzebny cząsteczce liganda na znalezienie na powierzchni receptora miejsca wiązania jest na ogół zbyt długi, aby możliwe było przeprowadzenie symulacji tego procesu za pomocą Dynamiki Molekularnej. Gdy jednak dostępne są fragmentaryczne dane na temat lokalizacji miejsca aktywnego, symulacje MD można z powodzeniem zastosować zarówno do przewidywania struktury kompleksów (Dominguez et al. 2003) jak i badania mechanizmów oddziaływań białek (Chen 2009). Dane o miejscu wiązania mogą pochodzić z różnorakich eksperymentów, ale dostępnych jest również wiele skutecznych metod ich przewidywania (de Vries & Bonvin 2008). Odmiennym sposobem jest połączenie Dynamiki Molekualrnej z innymi metodami dokowania. Modele uzyskane na przykład za pomocą sztywnego dokowania z „miękką" funkcją potencjału poddaje się symulacji MD w celu zrelaksowania układu i usunięcia zawad strukturalnych (Alonso et al. 2006).

3.2.4 Inne metody dokowania

Istnieje grupa metod hybrydowych, które łączą niektóre cechy wielu wymienionych wyżej algorytmów. W pracy (Sherman & Day 2006) autorzy proponują iteracyjny cykl obejmujący sztywne dokowanie z „miękką" funkcją potencjału przeplatany relaksacją łańcucha głównego i grup bocznych, aż do momentu osiągnięcia przez układ zbieżności. Podobne podejście zaprezentowano w pracy (Cavasotto & Abagyan 2004).

Strukturę kompleksu białkowego można w niektórych przypadkach uzyskać w procedurze analogicznej do modelowania homologicznego, stosowanego w predykcji struktur pojedynczych białek. M-TASSER (Chen & Skolnick 2008) to metoda oparta na modelu (Kihara & Lu 2001; Zhang et al. 2003; Zhang 2008), który bardzo dobrze sprawdził się w kolejnych edycjach konkursu CASP.

Na szczególną uwagę zasługuje metoda RosettaDock (Wang et al. 2007). Pozwala ona na modelowanie kompleksów białkowych z zachowaniem giętkości zarówno łańcucha głównego jak i grup bocznych białka. Metoda startuje z uzyskanych eksperymentalnie, bądź wygenerowanych przybliżonych struktur kompleksu, a następnie wykorzystuje sprawdzony algorytm modelu Rosetta (Rohl et al. 2004) do poszukiwania optymalnej konformacji

cząsteczek. Za pomocą RosettaDock przeprowadzono interesujące symulacje zwijania się białka w trakcie formowania symetrycznego homodimeru (Das et al. 2009).

Opisane powyżej metody posiadają swoje wady i zalety. Dokowanie do zbioru konformacji wydaje się być najbliższe postulowanemu mechanizmowi konformacyjnej selekcji (Boehr et al. 2009), lecz nie pozwala na modelowanie ścieżki tworzenia kompleksów. Z drugiej strony dynamika molekularna opisuje w szczegółowy sposób detale strukturalne, lecz zawodzi w przypadku, gdy utworzenie kompleksu wymaga znacznych zmian konformacyjnych. Jak pokazują wyniki kolejnych edycji eksperymentu CAPRI (Janin et al. 2003; Méndez et al. 2005; Wodak 2007; Fernández-Recio & Sternberg 2010) na dzień dzisiejszy nie da się wskazać jednej, najlepszej metody modelowania kompleksów białkowych, a jedynie grupę narzędzi, które działają relatywnie dobrze w pewnych szczególnych przypadkach. W najbliższym czasie wysiłki badaczy powinny skoncentrować się na zagadnieniach, z którymi obecnie nie radzi sobie żaden model: na modelowaniu układów białko/kwas nukleinowy i symulacjach równoczesnego zwijania się i łączenia białek w kompleksy.

III CABSDock – opracowanie automatycznej procedury dokowania białko-białko w zredukowanej przestrzeni konformacyjnej

Celem pracy doktorskiej było stworzenie automatycznej procedury dokowania białko-białko, która pozwalałaby na symulowanie procesu tworzenia się kompleksów białkowych z zachowaniem pełnej giętkości oddziałujących molekuł. W pierwotnym założeniu symulacje z użyciem modelu CABS miały stanowić jedyny etap, na którym molekuły podlegają znacznym zmianom konformacyjnym. Zastosowanie takiego podejścia do dokowania białko-peptyd zaowocowało obiecującymi wynikami (Kurcinski & Kolinski 2007b; Kurcinski & Kolinski 2007a; Kurcinski & Kolinski 2010; Horwacik et al. 2011), jednak wraz ze wzrostem rozmiaru cząsteczki liganda okazało się, że efektywne próbkowanie przestrzeni konformacyjnej typowych układów białko-białko wykracza poza możliwości nawet tak efektywnej metody próbkowania przestrzeni konformacyjnej jak REMC[14], na której opiera się szereg zastosowań modelu CABS. Konieczne zatem stało się zastosowanie jakiejś techniki wstępnego dokowania, która dostarczałaby przybliżonych modeli o niskiej rozdzielczości, następnie poprawianych za pomocą modelu CABS. W końcowym kształcie opisywana metoda urosła do wielostopniowej, automatycznej procedury, w której można odnaleźć elementy opisanych w poprzedniej części pracy typowych algorytmów dokowania białko-białko. Tym, co decyduje o unikalnym charakterze CABSDock jest zastosowanie modelu CABS w roli zarówno narzędzia służącego do generowania jak i oceniania oraz daleko posuniętego udokładniania struktur. Ostatnio ukazała się praca (Vajda et al. 2013), w której autorzy wykazują przewagę takich metod, w których próbkowanie i ocena realizowane są jednocześnie nad metodami, które rozbijają te zadania na dwa procesy, realizowane często za pomocą różnych modeli.

[14] Replica Exchange Monte Carlo

1 Ogólny schemat procedury dokowania

Dokowanie za pomocą CABSDock składa się z czterech głównych etapów, które częściowo pokrywają się z tymi opisanymi w sekcji II.3.2 o dokowaniu giętkim. Jedynymi danymi wejściowymi są pełno-atomowe, trójwymiarowe struktury dokowanych komponentów. Większe białko zostaje oznaczone jako receptor. Mniejsze traktuje się jako ligand, przy czym każdy z komponentów może sam być już kompleksem – tj. składać się z wielu łańcuchów białkowych. W przypadku przewidywania struktury homomeru, gdy komponenty kompleksu są identyczne, wciąż konieczne jest rozróżnienie na receptor i ligand, w celu zachowania zgodności formatów danych.

W pierwszym kroku wykonywana jest procedura sztywnego dokowania za pomocą publicznie dostępnego programu FTDOCK (Gabb et al. 1997). W wyniku działania programu generowanych jest 10000 zgrubnych struktur kompleksów, których komponenty znajdują się w konformacjach identycznych jak w stanie izolowanym. Modele różnią się jedynie wzajemnym ustawieniem receptora i liganda.

W następnym etapie, spośród 10000 modeli wybieranych jest 20, które poddane zostaną kolejnym symulacjom. Selekcja jest dokonywana na podstawie rankingu modeli opartego o specjalnie zaprojektowaną funkcję oceny przybliżającą zmianę energii swobodnej układu ΔF. Energia wewnętrzna obliczana jest za pomocą programu CABScore, który został opracowany w oparciu o potencjały statystyczne pola siłowego modelu CABS, natomiast wkład entropii do ΔF przybliżany jest w procesie analizy skupień wszystkich 10000 modeli.

Procedura sztywnego dokowania, wraz z analizą skupień i wyborem modeli do giętkiego dokowania została zoptymalizowana na testowym zestawie 160 struktur z bazy danych ZDOCK Benchmark (Chen et al. 2003; Mintseris et al. 2005). Parametry programów FTDOCK, CABScore oraz Clust z pakietu Bioshell (za pomocą którego przeprowadzono analizę skupień) zoptymalizowano tak, aby wśród 20 struktur wybieranych do etapu giętkiego dokowania znalazła się co najmniej jedna, o strukturze zbliżonej do natywnej.

Kolejny etap obejmuje symulację za pomocą modelu CABS. 20 struktur wybranych w poprzednim kroku poddawanych jest redukcji reprezentacji z pełno-atomowej do zawierającej położenia jedynie atomów węgla w pozycji α oraz

zrzutowanych na siatkę sześcienną modelu CABS. Tak przygotowane struktury stanowią dane wejściowe dla algorytmu realizującego dynamikę Monte Carlo w zredukowanej przestrzeni konformacyjnej, przy zachowaniu całkowitej giętkości modelowanych molekuł. W wyniku działania programu CABS generowana jest kolekcja 1000 modeli, będąca zapisem przebiegu symulacji. Modele ograniczone są jedynie do atomów węgla w pozycji α. Oczywiście można było stosować większe lub mniejsze liczby struktur i modeli. Wybrane wartości stanowią kompromis pomiędzy dokładnością modelu a jego kosztami obliczeniowymi.

Na ostatni etap opisywanej procedury składają się cztery procesy mające na celu wybór pięciu[15] końcowych modeli w pełno-atomowej reprezentacji. W pierwszej kolejności, za pomocą analizy skupień następuje selekcja pięciu struktur spośród 1000 wygenerowanych przez program CABS. Następnie, wybrane modele są rekonstruowane z uproszczonej do pełno-atomowej reprezentacji za pomocą programów BBQ (Gront et al. 2007) i SCWRL (Krivov et al. 2009). Tak przygotowane struktury są poddawane procedurze udokładniania za pomocą minimalizacji energii w empirycznym polu siłowym za pomocą pakietu Gromacs (Pronk et al. 2013). Wartość energii po minimalizacji stanowi podstawę do wybrania pięciu końcowych modeli.

Cała procedura została zautomatyzowana za pomocą dodatkowych programów i skryptów tak, aby kolejne etapy wykonywały się bez potrzeby jakiejkolwiek ingerencji użytkownika. Jednocześnie segmentowa konstrukcja CABSDock pozwala na łatwą modyfikację programów realizujących poszczególne etapy modelowania. Na chwilę obecną CABSDock został zaimplementowany na klastrze obliczeniowym Ośrodka Komputerowego Wydziału Chemii UW. Pod adresem www.biocomp.chem.uw.edu.pl/tools/cabsdock udostępniono publicznie program CABS, zmodyfikowany do symulowania białek wielołańcuchowych. W kolejnych sekcjach znajdują się szczegółowe opisy poszczególnych etapów procedury.

[15] pakiet CABSDock był projektowany między innymi z myślą o udziale w zawodach CAPRI, gdzie wymagane jest dostarczenie właśnie 5 modeli.

2 Dokowanie sztywne – FTDOCK

FTDOCK to program realizujący sztywne dokowanie białko-białko. Został opracowany w grupie Paula Bates'a z Biomolecular Modelling Laboratory w Imperial Cancer Research Fund w Wielkiej Brytanii. Jest rozpowszechniany na zasadach licencji GNU General Public Licence, przez co jest on dostępny za darmo zarówno dla użytkowników akademickich jak i komercyjnych. W programie FTDOCK zaimplementowano algorytm autorstwa Katchalski-Katzira (Katchalski-Katzir et al. 1992) wykorzystujący Szybką Transformację Fouriera (FFT) do wyczerpującego przeszukiwania sześciowymiarowej przestrzeni konformacyjnej translacji i rotacji cząsteczki liganda względem cząsteczki receptora. Poniżej opisano najważniejsze elementy algorytmu.

2.1 Reprezentacja molekuł

Struktury receptora i liganda rzutowane są na dwie identyczne siatki sześcienne, o stałych siatkowych równych D, według następującej procedury: Każdemu węzłowi siatki przypisywana jest wartość „1", jeżeli w odległości nie większej niż R znajduje się co najmniej jeden atom lub wartość „0" w przeciwnym wypadku. Wartość R jest rzędu atomowego promienia van der Waalsa. Dodatkowo dla węzłów siatki związanej z receptorem, którym przypisano wartość „1" oraz których odległość do najbliższego węzła o wartości „0" jest większa niż parametr L (grubość warstwy powierzchniowej), przypisywana jest ujemna wartość S. Węzły te reprezentują wewnętrzny rdzeń molekuły receptora, niezaangażowany w wiązanie liganda.

2.2 Funkcja oceny

FTDOCK ocenia generowane struktury kompleksów tylko na podstawie komplementarności kształtów oddziałujących powierzchni. Niech $a(i,j,k)$ i $b(i,j,k)$ będą dyskretnymi funkcjami o wartościach odpowiadających węzłom o współrzędnych (i,j,k) siatek odpowiednio receptora i liganda. Wtedy wartość funkcji oceny $E(x,y,z)$ wynosi:

$$E(x,y,z) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} a(i,j,k) \cdot b(i+x, j+y, k+z) \qquad [3.1]$$

gdzie *N* to liczba węzłów siatki wzdłuż jednego kierunku, a *V(x,y,z)* to wektor przesunięcia siatki liganda względem siatki receptora. *a(i,j,k)* może przyjmować wartości 0,1lub *S*, gdzie *S* jest ujemne. *b(i,j,k)* przyjmuje jedynie wartości 0 lub 1.

Iloczyn *a·b* ma wartość niezerową jedynie w dwóch przypadkach:

- gdy *a=1* i *b=1*, wtedy *a·b=1* – oznacza korzystny (zwiększający wartość *E*) kontakt pomiędzy węzłem na powierzchni receptora i węzłem liganda,
- gdy *a=S* i *b=1*, to *a·b=S* – oznacza niekorzystny (zmniejszający wartość *E*) kontakt węzła liganda z węzłem wewnętrznego rdzenia receptora – zawadę steryczną.

2.3 Przeszukiwanie przestrzeni konformacyjnej

Przestrzeń konformacyjna układu posiada sześć wymiarów: trzy odpowiedzialne za rotacje i trzy za translacje liganda względem receptora.

2.3.1 Skanowanie rotacji

Niech α, β, γ oznaczają obrót struktury liganda wokół prostopadłych osi układu współrzędnych związanego z siatką receptora. Dla wszystkich kombinacji, zmieniających się skokowo o parametr θ, zmiennych α, β i γ, struktura liganda obracana jest o kąty α, β, γ wokół odpowiednich osi, a następnie rzutowana na siatkę, po czym sprawdzane są wszystkie możliwe translacje siatek. Dla danej rotacji zachowywane są jedynie 3 najwyżej ocenione struktury. Po sprawdzeniu wszystkich rotacji, spośród $3 \cdot \left(\frac{\pi}{\theta}\right)^3$ zachowanych struktur, wypisywanych jest 10000 o najwyższych ocenach.

2.3.2 Skanowanie translacji

Funkcja *E(x,y,z)* w definicji z równania [3.1] jest splotem[16] funkcji *a(x,y,z)* i *b(x,y,z)*. Bezpośrednie obliczanie wartości funkcji *E(x,y,z)* dla konkretnych wartości x_i,y_i,z_i wymaga N^3 operacji mnożenia oraz dodawania. Liczba kombinacji liczb *x,y,z* wynosi również N^3, zatem obliczenie funkcji *E(x,y,z)* dla wszystkich możliwych translacji siatki liganda względem receptora wiąże się z kosztem obliczeniowym rzędu $o(N^6)$. Na mocy twierdzenia Borela o splocie transformata Fouriera splotu funkcji *f* i *g* równa jest iloczynowi transformat Fouriera tych funkcji:

$$\mathcal{F}(f * g) = \mathcal{F}(f) \cdot \mathcal{F}(g) \quad \Rightarrow \quad f * g = \mathcal{F}^{-1}(\mathcal{F}(f) \cdot \mathcal{F}(g)) \qquad [3.2]$$

Z równań [3.1] i [3.2] otrzymujemy:

$$E(x,y,z) = a(x,y,z) * b(x,y,z) = \mathcal{F}^{-1}(\mathcal{F}(a(x,y,z)) \cdot \mathcal{F}(b(x,y,z))) \qquad [3.3]$$

Dzięki wysoce wydajnemu algorytmowi FFT obliczanie powyższego wyrażenia charakteryzuje się kosztem obliczeniowym rzędu $o(N^3 log N^3)$, co dla dużych wartości N przekłada się na istotne skrócenie czasu wykonywania w stosunku do równania [3.1], którego obliczanie wymaga N^6 operacji.

3 Selekcja modeli do dokowania giętkiego

W wyniku działania programu FTDOCK generowanych jest 10000 struktur kompleksu, spośród których wybieranych jest 20 najlepiej ocenionych, które posłużą, jako dane wejściowe do dokowania giętkiego. Selekcja dokonywana jest na podstawie rankingu sporządzonego za pomocą specjalnie skonstruowanej funkcji oceny, przybliżającej zmianę energii swobodnej ΔF pomiędzy stanem związanym a swobodnym. Energia swobodna jest termodynamiczną funkcją stanu równą z definicji:

[16] splot funkcji *f(x)* i *g(x)* całkowalnych w przedziale (-∞,+∞) jest działaniem, którego wynikiem jest funkcja *h(x)* równa z definicji:

$$h(x) = f(x) * g(x) = \int_{-\infty}^{+\infty} f(t)g(x-t)dt \qquad [3.4]$$

$$F = U - TS \qquad [3.5]$$

gdzie U to energia wewnętrzna, T – temperatura, a S – entropia.

Energia wewnętrzna to termodynamiczna funkcja stanu określająca całkowity zasób energii układu. Równa jest sumie energii oddziaływań międzycząsteczkowych i wewnątrz-cząsteczkowych, a także energii ruchu cieplnego cząsteczek oraz wszystkich innych rodzajów energii występujących w układzie. W stanie związanym struktury poszczególnych komponentów kompleksów wygenerowanych w trakcie dokowania sztywnego są identyczne jak w stanie niezwiązanym – ich energia wewnętrzna nie zmienia się. Zatem zmiana energii wewnętrznej ΔU całego kompleksu równa jest energii oddziaływania pomiędzy komponentami E_{inter}. W celu wyznaczenia wartości E_{inter} opracowano program CABScore, w którym zaimplementowano jedynie te elementy pola siłowego modelu CABS, które są odpowiedzialne za oddziaływanie pomiędzy łańcuchami białkowymi.

Entropia jest termodynamiczną funkcją stanu, określającą kierunek przebiegu reakcji spontanicznych w układzie izolowanym. Termodynamika definiuje entropię, jako:

$$S = k_B ln\,(\Omega) \qquad [3.6]$$

gdzie k_B – stała Boltzmanna, a Ω to liczba sposobów, na jakie może być realizowany dany stan makroskopowy. W praktyce wyznaczenia wartości Ω jest niemożliwe dla tak dużych układów. Ponadto, aby zastosować równanie [3.5] konieczne by było również powiązanie entropii z bezwymiarowymi wartościami energii i temperatury, jakie są używane w modelu CABS. Aby jednak uwzględnić czynnik entropowy w rankingu wygenerowanych w trakcie dokowania sztywnego modeli przeprowadzono analizę skupień. Struktury zostały pogrupowane w zbiory według wzajemnego podobieństwa wyrażonego jako RMSD.

3.1 Obliczanie energii oddziaływania E_{inter} – CABScore

Energia oddziaływania w programie CABScore obliczana jest według następującego wyrażenia:

$$E_{inter} = \sum_{m=1}^{N} \sum_{n=m+1}^{N} \sum_{i=1}^{L_m} \sum_{j=1}^{L_n} E_{resid}(i,j) \qquad [3.7]$$

gdzie N – liczba łańcuchów białkowych w kompleksie, L_m – liczba reszt w m-tym łańcuchu, natomiast $E_{resid}(i,j)$ to dwuargumentowa funkcja zwracająca wartość energii oddziaływania pomiędzy dwoma resztami białkowymi. Funkcja E_{resid} jest zdefiniowana w następujący sposób[17]:

$$E_{inter} = \begin{cases} 0 & d_{ij} > D_{max}(\dots) \\ E(\dots) - E_0 \quad gdy & D_{min}(\dots) \le d_{ij} \le D_{max}(\dots) \\ E_{rep} & d_{ij} < D_{min}(\dots) \end{cases} \qquad [3.8]$$

- A_i – rodzaj i-tego aminokwasu;
- S_i – lokalna konformacja i-tego aminokwasu, może przyjmować dwie wartości w zależności od odległości $R^{13}_i = |\mathbf{R}^{CA}_{i+1} - \mathbf{R}^{CA}_{i-1}|$, gdzie $\mathbf{R}^{CA}_i$ to wektor położenia i-tego atomu węgla w pozycji α. Jeżeli $R^{13}_i < 6.0$ Å, to S_i oznacza konformacje zamkniętą (charakterystyczną dla α-helis), w przeciwnym wypadku – otwartą (β-kartek);
- F_{ij} – orientacja kontaktujących się reszt, przyjmuje trzy wartości: równolegle (P), antyrównolegle (A) oraz pośrednio (M) w zależności od wartości iloczynu skalarnego wektorów $\mathbf{b}_i$ i $\mathbf{b}_j$, zdefiniowanych jako $\mathbf{b}_i = (\mathbf{V}_{i-1} - \mathbf{V}_i) / |\mathbf{V}_{i-1} - \mathbf{V}_i|$, gdzie $\mathbf{V}_i = \mathbf{R}^{CA}_{i+1} - \mathbf{R}^{CA}_i$, a $\mathbf{R}^{CA}_i$ to wektor położenia i-tego atomu węgla w pozycji α. Jeżeli $\mathbf{b}_i \cdot \mathbf{b}_j > 0.5$, to F_{ij} =P lub jeżeli $\mathbf{b}_i \cdot \mathbf{b}_j < 0.5$, to F_{ij} =A lub F_{ij} =M gdy $-0.5 \le \mathbf{b}_i \cdot \mathbf{b}_j \le 0.5$;
- d_{ij} – długość wektora $\mathbf{CM}_i - \mathbf{CM}_j$, gdzie $\mathbf{CM}_i$ – środek geometryczny wszystkich ciężkich atomów łańcucha bocznego wraz z atomem węgla w pozycji α i-tego aminokwasu;
- $D_{max}(\dots) = D(\dots)$ jeżeli $E(\dots) \ge 0$ lub
- $D_{max}(\dots) = D(\dots) + D_1$ jeżeli $E(\dots) < 0$;
- $D_{min}(\dots) = \max[D(\dots) - D_2, 2.0\text{ Å}]$;
- E_0, D_1, D_2 – parametry, których wartości były optymalizowane;

[17] dla przejrzystości w zamieszczonych wzorach zastosowano podstawienie '…' ≡ $A_i, A_j, S_i, S_j, F_{ij}$

- *E(...), D(...)* stabelaryzowane parametry potencjału par pola siłowego CABS, wyprowadzone za pomocą przybliżenia quasi-chemicznego, na zbiorze unikalnych powierzchni oddziaływań międzybiałkowych PIBASE (Davis & Sali 2005).

Potencjał par zdefiniowany w powyższy sposób wyróżnia się spośród innych potencjałów tego typu uwzględnieniem szczegółowego kontekstu, w jakim dochodzi do kontaktu. Znakomita większość statystycznych potencjałów par uzależnia wartość energii oddziaływania jedynie od rodzaju i odległości pomiędzy resztami, podczas gdy zastosowany tutaj bardzo szczegółowy opis interakcji, pozwala na rozróżnienie charakteru oddziaływania wynikającego jedynie z różnicy w lokalnej konformacji łańcucha białkowego.

3.2 Analiza skupień

W celu oszacowania wartości entropii każdego spośród 10000 modeli wygenerowanych przez program FTDOCK, wykonywana jest analiza skupień. W pierwszej kolejności generowana jest macierz podobieństwa pomiędzy modelami, gdzie miarą podobieństwa jest wartość RMSD. Następnie w procedurze klasteryzacji hierarchicznej modele grupowane są w zbiory według wzajemnego podobieństwa. Pomimo iż modele przynależące do jednego zbioru potrafią wykazywać czasem dosyć istotne różnice strukturalne, założono, że reprezentują one ten sam stan energetyczny. Jako struktura reprezentatywna z każdego zbioru wybierana była ta, dla której wartość energii wewnętrznej wyznaczonej za pomocą programu CABScore była najniższa.

3.3 Ranking modeli

Z 20 największych klastrów wybierane są struktury o najniższej energii wyliczonej za pomocą CABScore. Struktury te stanowią dane wejściowe dla dokowania giętkiego, realizowanego w kolejnym kroku za pomocą programu CABS.

4 Dokowanie giętkie za pomocą modelu CABS

4.1 Opis modelu CABS

CABS to model służący do badania dynamiki i termodynamiki białek, opracowany w 2004 roku przez Kolińskiego (Kolinski 2004; Kolinski et al. 2006). Pierwotnie został skonstruowany do przewidywania struktur białek, lecz z czasem okazał się bardzo użyteczny w innych zastosowaniach, takich jak rekonstrukcja fragmentów struktur białkowych, badanie dynamiki stanu natywnego białek, badanie mechanizmów zwijania się białek, czy wreszcie modelowanie oddziaływań białkowych. Model został już wcześniej szczegółowo opisany (Kolinski 2004), poniżej przedstawiono jego najważniejsze założenia.

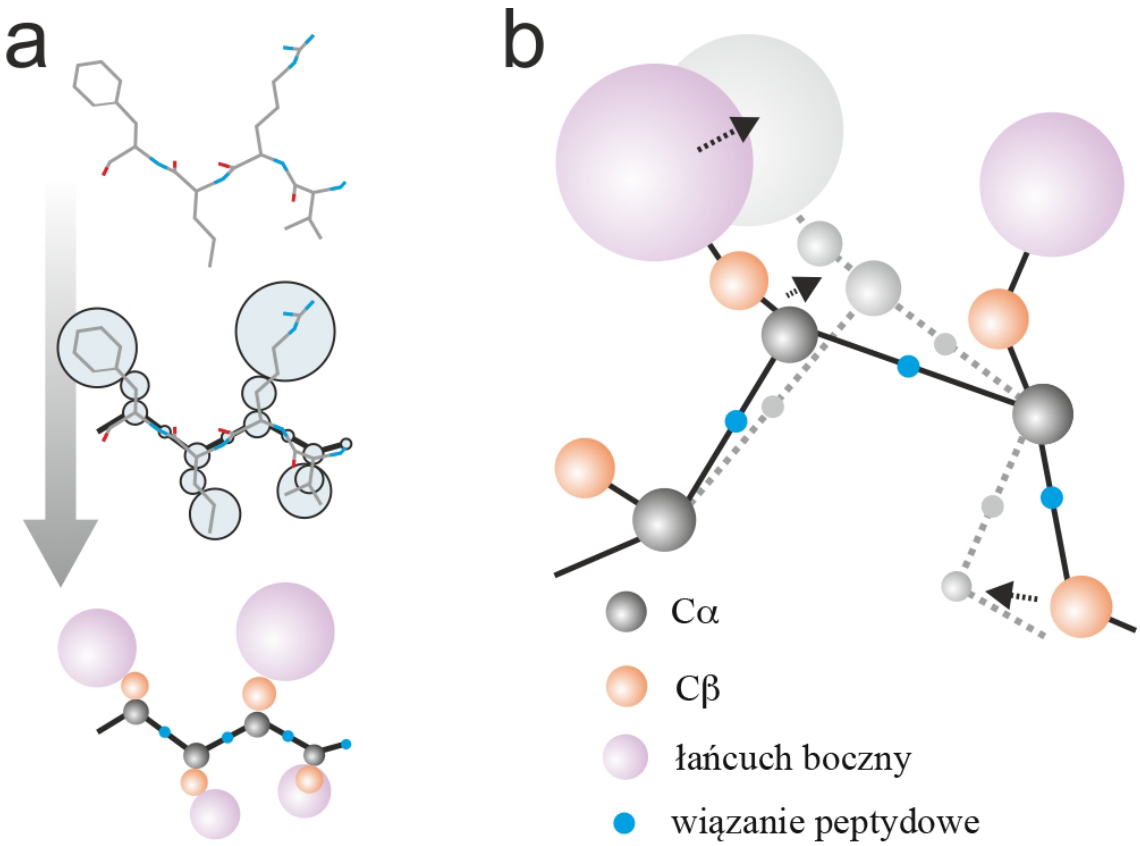
4.1.1 Reprezentacja molekuł

CABS to skrót pochodzący od pierwszych liter nazw pseudoatomów wybranych do zredukowanej reprezentacji reszt aminokwasów: **CA** (atom węgla w pozycji α), C**B** (atom węgla w pozycji β) oraz **S**G (zjednoczony pseudoatom grupy bocznej[18]). Dodatkowo na potrzeby definicji wiązań wodorowych pomiędzy atomami łańcucha głównego, definiowany jest również pseudoatom wiązania peptydowego, zlokalizowany w środku geometrycznym pseudowiązania Cα-Cα. Zastosowanie jedynie tych czterech centrów oddziaływań, w miejsce od 7 (Glicyna) do 24 (Tryptofan) atomów w przypadku reprezentacji pełnoatomowej, pozwala na bardzo istotne skrócenie czasu obliczeń.

Dodatkowo model CABS wykorzystuje sześcienną siatkę, do której węzłów ograniczone są położenia atomów Cα. Położenia pozostałych atomów są wyliczane na podstawie wektorów łączących atom Cα danej reszty z atomami Cα reszty poprzedzającej i następującej po niej w łańcuchu polipeptydowym. Stała siatkowa w modelu CABS została arbitralnie ustalona na 0.61Å. Zastosowano również ograniczenie dopuszczalnej fluktuacji długości wektorów Cα-Cα oraz kątów pomiędzy nimi, tak aby odpowiadały rzeczywistym wartościom spotykanym

[18] (ang.) side group

w białkach. Przy takiej definicji siatki przestrzennej uzyskano jedynie 800 możliwych wektorów łączących dwa dowolne atomy Cα następujące po sobie w łańcuchu białkowym. Jest to liczba wystarczająco duża, aby rzutując strukturę na siatkę nie wprowadzać istotnego błędu[19], a jednocześnie na tyle mała, że możliwe jest przeliczenie, stabelaryzowanie i zapisanie w pamięci operacyjnej wielu wielkości zależnych od wektorów Cα-Cα (np. iloczynów skalarnych pomiędzy wektorami, wektorów położeń atomów Cβ względem atomów Cα itp.), tak aby w trakcie symulacji nie trzeba było ich obliczać, ale jedynie szybko odczytywać.



Rys. 3.1

Zredukowana reprezentacja cząsteczki białka w modelu CABS (a) oraz przykład mikromodyfikacji zmieniającej konformację pojedynczej reszty (b).

[19] średnia wartość RMSD pomiędzy strukturą rzeczywistą, a zrzutowaną na siatkę CABS'a wynosi 0.35Å i jest mniejsza od dokładności z jaką wyznaczono eksperymentalnie znakomitą większość struktur zdeponowanych w bazie PDB

4.1.2 Próbkowanie przestrzeni konformacyjnej

Próbkowanie przestrzeni konformacyjnej w modelu CABS realizowane jest za pomocą czterech typów losowych mikromodyfikacji, którym poddawane są jedynie atomy Cα. Pozostałe atomy „podążają" za ruchem łańcucha głównego – ich położenia są odczytywane z tabel na podstawie położenia atomów Cα. Mikromodyfikacje występują jedynie lokalnie, lecz w skali czasowej typowej symulacji składają się na globalne ruchy całego układu.

Atomy poruszane są losowo zgodnie z zasadami dynamiki Monte Carlo (Metropolis 1987), tzn. o zaakceptowaniu nowej konformacji decyduje warunek Metropolisa (Metropolis et al. 1953): nowa konformacja jest zawsze akceptowana jeśli prowadzi do obniżenia energii układu, w przeciwnym przypadku prawdopodobieństwo jej zaakceptowania maleje wykładniczo ze wzrostem energii. Jest to tzw. „asymetryczny schemat Metropolisa" najczęściej wykorzystywany w algorytmach dynamiki MC.

Dodatkowo w celu zwiększenia efektywności próbkowania w modelu CABS zaadoptowano metodę wymiany replik (Swendsen & Wang 1986). Symulacja dynamiki prowadzona jest w kilku kopiach różniących się temperaturą. Co pewien czas algorytm wymienia temperatury pomiędzy dwoma kopiami z prawdopodobieństwem proporcjonalnym wykładniczo do różnicy energii i odwrotnie wykładniczo do różnicy temperatur pomiędzy nimi. Dzięki temu unika się sytuacji, w której symulacja zbiega do lokalnego minimum na powierzchni energii potencjalnej, do czego często mogą prowadzić symulacje klasycznej dynamiki Monte Carlo.

4.1.3 Pole siłowe

Pole siłowe modelu CABS ma prawie wyłącznie charakter statystyczny, tzn. wartości poszczególnych typów oddziaływań nie są obliczane na podstawie wzorów mających podstawy fizyczne, ale odczytywane z tabel i histogramów sporządzonych na podstawie analizy statystycznej regularności zaobserwowanych w strukturach rozwiązanych eksperymentalnie i zdeponowanych w bazie PDB. W skład energii całkowitej wchodzą następujące człony:

- **oddziaływania bliskiego zasięgu[20] niezależne od sekwencji** – jest to zestaw warunków narzuconych na wektory Cα-Cα mających na celu nadanie łańcuchowi głównemu sztywności charakterystycznej dla białek (np. odległość pomiędzy kolejnymi atomami Cα musi zawierać się pomiędzy 3.28Å, a 4.27Å).
- **oddziaływania bliskiego zasięgu zależne od sekwencji** – potencjał odpowiedzialny za lokalną konformację krótkich (obejmujących 3-5 reszt) fragmentów w zależności od rodzaju reszt z których się składają. Potencjał ten ma postać histogramów odległości pomiędzy resztami oddalonymi od siebie o 2, 3 i 4 pozycje w łańcuchu głównym, uwzględniających chiralność i lokalną strukturę drugorzędową. Zysk energetyczny jest proporcjonalny wykładniczo do częstości występowania zaobserwowanej odległości w strukturach wyznaczonych eksperymentalnie, które posłużyły do konstrukcji histogramów.
- **oddziaływania dalekiego zasięgu[21] niezależne od sekwencji** – potencjały tego typu obejmują człony odpowiedzialne za odpychanie się atomów na bliskich odległościach (wyłączona objętość), jak również za tworzenie się wiązań wodorowych pomiędzy atomami łańcucha głównego.
- **oddziaływania dalekiego zasięgu zależne od sekwencji** – dwuciałowy potencjał kontaktowy odpowiedzialny za oddziaływania grup bocznych, mający postać tabel, zależnych od rodzaju kontaktujących się reszt, ich przestrzennej orientacji i lokalnych konformacji w jakich się znajdują. Tabele zawierają minimalne i maksymalne odległości na jakich dany kontakt może wystąpić oraz zysk energetyczny wynikający z zaistnienia kontaktu. Tabele zostały skonstruowane na podstawie analizy struktur z bazy PDB. Potencjał ten niejawnie obejmuje również oddziaływania elektrostatyczne pomiędzy naładowanymi resztami, wiązania wodorowe pomiędzy grupami bocznymi, mostki siarczkowe oraz wpływ rozpuszczalnika (efekt hydrofobowy).
- **potencjał centrosymetryczny** – potencjał odpowiedzialny za zbliżony do kulistego kształt jaki najczęściej przyjmują białka globularne.

[20] pomiędzy resztami sąsiadującymi w sekwencji
[21] pomiędzy resztami oddalonymi od siebie w sekwencji

4.1.4 Adaptacja modelu CABS do układów wielołańcuchowych

Model CABS został w pierwotnej wersji zaprojektowany do symulowania dynamiki pojedynczych łańcuchów białkowych. Dyskretna przestrzeń konformacyjna oraz statystyczne pole siłowe pozwoliło na wykorzystanie wielu technicznych zabiegów mających na celu przyspieszenie obliczeń, w szczególności na obliczenie wielu potrzebnych w trakcie symulacji wielkości i zapisanie ich w pamięci operacyjnej komputera.

W celu umożliwienia symulacji dynamiki wielu łańcuchów, CABS został poddany kilku modyfikacjom, przy czym zadbano, aby nie wpłynęły one na wzrost kosztu obliczeniowego innego niż wynikający z liczby atomów w symulowanym układzie. Próbkowanie odbywa się zatem „w turach" – w momencie gdy modyfikowana jest konformacja jednego łańcucha, reszta układu jest zamrożona. Oddziaływania bliskiego zasięgu (sąsiadujące w sekwencji) rozciągają się w oczywisty sposób jedynie pomiędzy resztami tego samego łańcucha, podczas gdy oddziaływania dalekiego zasięgu obejmują kontakty pomiędzy resztami zarówno tego samego, jak i różnych łańcuchów. Wartość potencjału centrosymetrycznego obliczana jest dla każdego łańcucha z osobna.

Model CABS dzięki zastosowaniu zredukowanej reprezentacji molekuł, dyskretnej przestrzeni konformacyjnej, metody wymiany replik oraz statystycznym potencjałom pola siłowego pozwala na symulowanie procesów nieosiągalnych dla metod opartych na dynamice molekularnej, zarówno ze względu na rozmiar symulowanych układów, jak i czas ich trwania: CABS był wykorzystywany w modelowaniu struktury ludzkiej Telomerazy (Steczkiewicz et al. 2011) – kompleksu składającego się z ponad 1000 aminokwasów, jak również w modelowaniu formowania się włókien amyloidowych (Małolepsza et al. 2005), co *in vivo* zachodzi w czasie od minut do wielu dni, a nawet lat.

4.1.5 CABS-flex: implementacja mechanizmu „konformacyjnej selekcji"

Model CABS pozwala na symulowanie dynamiki białek ze zmiennym stopniem giętkości molekuł, poprzez narzucenie na modelowany układ więzów odległości na wybrane atomy Cα. Wybór w którym miejscu i jak silne zastosować

więzy jest zadaniem nietrywialnym i mającym duże znaczenie na końcowy wynik symulacji, zwłaszcza w przypadku modelowania układów wielu białek. Zastosowanie zbyt wielu lub zbyt silnych więzów prowadzi do symulacji, w których niewiele się dzieje – molekuły są zbyt sztywne, aby efektywnie dopasować do siebie swoje konformacje. Z drugiej strony pozostawienie cząsteczek zbyt giętkimi prowadzić może do ich rozwinięcia.

W pracy z 2008 roku (Dobbins et al. 2008) autorzy badali 20 białek, które podlegają znacznym zmianom konformacyjnym podczas łączenia się w kompleksy. Okazało się, że analiza drgań normalnych konformacji natywnych w stanie niezwiązanym pozwala przewidywać kierunek i amplitudę zmian konformacyjnych podczas formowania kompleksów. Odkrycie to potwierdza postulowany w 2009 roku mechanizm „konformacyjnej selekcji" (Boehr et al. 2009), zgodnie z którym izolowane komponenty kompleksów białkowych obsadzają jednocześnie wiele konformacji, z których dopiero te najkorzystniejsze energetycznie prowadzą do stanu związanego. W opublikowanej ostatnio pracy (Jamroz et al. 2013) pokazano, że dzięki specyficznie dobranym więzom odległości, model CABS pozwala na odtworzenie drgań normalnych stanu natywnego. Taki sam schemat więzów wykorzystano w procedurze CABSDock, co pozwoliło na modelowanie mechanizmu „konformacyjnej selekcji".

Więzy odległości na atomy Cα zostały wyselekcjonowane ze struktur natywnych izolowanych komponentów przy spełnionych następujących warunkach:

- maksymalna odległość pomiędzy związanymi atomami jest mniejsza od 8Å;
- związane są jedynie reszty, którym program DSSP[22] przypisał strukturę drugorzędową inną niż kłębek (C);
- reszty znajdujące się w tym samym łańcuchu białkowym są związane jeżeli odstęp pomiędzy nimi wynosi co najmniej 2 pozycje w sekwencji.

[22] DSSP to program, który na podstawie trójwymiarowych współrzędnych atomów przypisuje każdej z reszt łańcucha białkowego lokalną strukturę drugorzędową w jednoliterowym kodzie: α-helisa (H), 3_{10}-helisa (G), π-helisa (I), β-wstęga (E), β-mostek (B), β-zakręt (T), zakręt (S) i kłębek (C). Zaimplementowano w nim algorytm autorstwa W. Kabsh'a i C. Sander'a (Kabsch & Sander 1983).

5 Selekcja i wygładzanie końcowych modeli

W wyniku działania programu CABS otrzymuje się zapis przebiegu symulacji w postaci pliku zawierającego szereg trajektorii. W pliku tym zapisane są współrzędne wszystkich atomów symulowanego układu uporządkowane w kolejne klatki na podobieństwo klatek na taśmie filmowej. Częstość z jaką aktualna konformacja układu jest zapisywana w trajektorii jest parametrem programu i może być modyfikowana. Arbitralnie ustalono, że trajektoria zawierać powinna 1000 modeli, co jest liczbą wystarczająco dużą, aby uzyskać wiarygodne statystyki, a jednocześnie na tyle małą, że koszt obliczeniowy kolejnych etapów nie jest zbyt wysoki. Ponadto w trajektorii pominięto konformacje replik znajdujących się w wysokich temperaturach, pozostawiając jedynie tę, której temperatura była aktualnie najniższa.

5.1 Analiza skupień

W kolejnym kroku modelowania trajektoria wygenerowana przez CABSDock poddawana jest analizie skupień z wykorzystaniem hierarchicznej analizy skupień (klasteryzacji). Najpierw oblicza się macierz odległości pomiędzy modelami. Istnieje wiele miar podobieństwa strukturalnego, które mogą być wykorzystane do wygenerowania takiej macierzy. W tej pracy zdecydowano się na RMSD obliczane na całym modelowanym układzie. Algorytm klasteryzacji hierarchicznej wykonuje maksymalnie N (liczba modeli w klastrowanym zbiorze) kroków, w każdym scalając dwa najbardziej do siebie podobne struktury/podzbiory, aż do momentu gdy żadne dwa podzbiory nie różnią się od siebie bardziej, niż zadana z góry wartość maksymalnej odległości pozwalającej na scalenie ich razem. Jako strukturę reprezentatywną dla każdego z otrzymanych podzbiorów, wybiera się medoid, czyli tę, która jest najbardziej podobna do geometrycznego centroidu obliczonego dla całego podzbioru. Struktury reprezentatywne z pięciu największych klastrów są wybierane do następnego etapu modelowania.

5.2 Odbudowa modeli do pełnoatomowej reprezentacji

CABSDock generuje modele w reprezentacji zredukowanej jedynie do atomów Cα. Procedura odbudowy modeli do pełnoatomowej reprezentacji składa się z dwóch etapów.

5.2.1 Odbudowa łańcucha głównego

Atomy łańcucha głównego odbudowywane są za pomocą programu BBQ (Gront et al. 2007). Algorytm przeszukuje bazę tetrapeptydowych fragmentów struktur z bazy PDB dopasowując je do odbudowywanego łańcucha, a następnie kopiuje współrzędne atomów łańcucha głównego z dopasowanego do odbudowywanego fragmentu.

5.2.2 Odbudowa grup bocznych

Atomy grup bocznych odbudowywane są za pomocą programu SCWRL (Krivov et al. 2009). Program ten przeszukuje bibliotekę rotamerów zależnych od konformacji łańcucha głównego oraz przeprowadza minimalizację prostej funkcji energii. Potencjał ten zależny jest od częstości występowania danego rotameru w znanych strukturach białkowych oraz zawiera niespecyficzny człon odpowiadający za wyłączoną objętość.

5.3 Końcowe wygładzanie modeli

Po odbudowie do pełnoatomowej reprezentacji wszystkie pięć modeli poddawanych jest minimalizacji energii w empirycznym polu siłowym. Do tego zadania wykorzystano pakiet Gromacs (Pronk et al. 2013). Modele były poddawane krótkiej symulacji dynamiki molekularnej w polu siłowym AMBER99 (Wang et al. 2000) z jawnie występującymi molekułami rozpuszczalnika, a następnie minimalizacji energii metodą największego spadku. Końcowa energia całkowita po minimalizacji stanowi podstawę do końcowego rankingu modeli.

IV Modelowanie kompleksów białkowych – uzyskane wyniki

Rozdział ten zawiera skrócone omówienie wyników badań wykorzystujących opracowaną metodę modelowania kompleksów białek. Większość uzyskanych wyników zawarta jest w już opublikowanych pracach autora oraz jednej zgłoszonej do druku (Załączniki). Rozdział ten składa się z dwóch części: dokowanie białko-peptyd i białko-białko.

Modelowanie układu białko-peptyd jest podobne do modelowania układu biało-białko, różnica dotyczy jedynie wielkości dokowanych molekuł. Zwyczajowo peptydy dłuższe niż 50 aminokwasów uznaje się za białka. Największe białka jednodomenowe składają się z ponad 600 aminokwasów, przy czym znakomita większość zawiera się w przedziale 100-200 reszt. To arbitralne rozgraniczenie okazuje się dobrze sprawdzać w przypadku modelowania za pomocą modelu CABS – algorytm pozwala na bardzo efektywne przeszukiwanie przestrzeni konformacyjnej układu receptor-ligand, tak że cząsteczka liganda ma szanse zwiedzić wiele miejsc na powierzchni receptora w poszukiwaniu najkorzystniejszego energetycznie miejsca wiązania, o ile ligand (mniejsze białko) nie jest zbyt duży. Do modelowania układów, w których ligand składa się z więcej niż 50-60 reszt, konieczne jest zastosowanie wstępnego dokowania sztywnego.

1 Dokowanie białko-peptyd

1.1 Modelowanie trójwymiarowych struktur kompleksów receptorów jądrowych i peptydów imitujących czynniki transkrypcyjne

Receptory jądrowe stanowią ważną klasę białek występujących w jądrach komórek zwierzęcych. Między innymi pełnią one w organizmach funkcję czynników transkrypcyjnych: poprzez zdolność do łączenia się bezpośrednio z nicią DNA regulują ekspresję genów, a tym samym są odpowiedzialne za rozwój, metabolizm i homeostazę całego organizmu.

Receptory jądrowe łączą się ze specyficznymi hormonami takimi jak: witamina D, estrogen, tyroksyna i inne, w wyniku czego zmieniają swoją konformację, co powoduje z kolei zmianę ich aktywności transkrypcyjnej. Poza ligandami pierwszorzędowymi receptory jądrowe łączą się również z innymi białkami, bądź peptydami – koaktywatorami, korepresorami i różnymi czynnikami transkrypcyjnymi. Mnogość funkcji, jak również oddziałujących partnerów powoduje, że receptory jądrowe stanowią ważny cel dla racjonalnego projektowania leków.

W pracy P.I przeprowadzono eksperyment w pełni giętkiego dokowania krótkich peptydów – fragmentów czynników transkrypcyjnych do grupy 10 białek z rodziny receptorów jądrowych. W bazie PDB znajdują się ich struktury krystalograficzne wysokiej rozdzielczości, dzięki czemu możliwe było zweryfikowanie poprawności otrzymanych wyników.

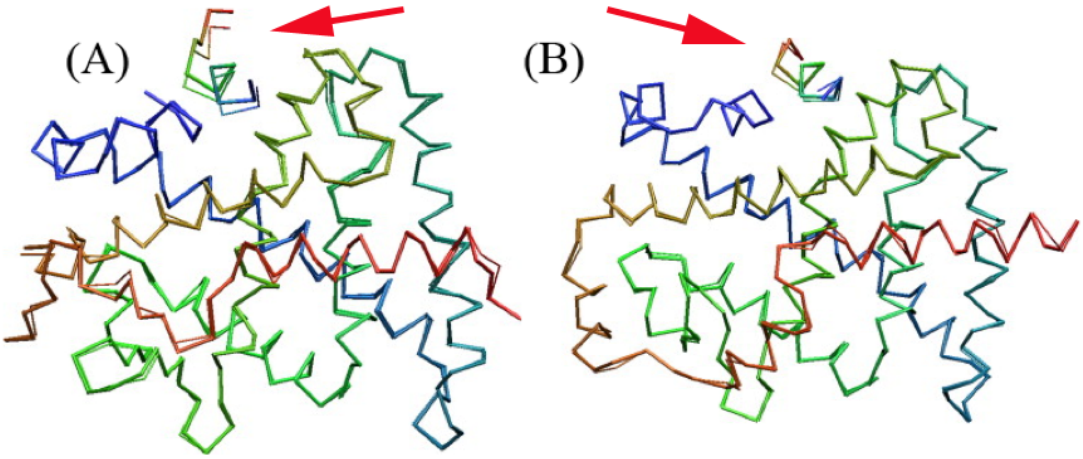
W symulacjach zastosowano następujące warunki początkowe:

- cząsteczka receptora znajdowała się w swojej natywnej konformacji,
- cząsteczka liganda znajdowała się w losowej konformacji i była umieszczona w losowym miejscu na sferze o promieniu równym 40 Å i środku zlokalizowanym w środku masy receptora,
- na cząsteczkę receptora nałożone zostały słabe więzy utrzymujące ją w konformacjach bliskich natywnej.

Zastosowano następujący schemat wyboru końcowych modeli: trajektorie będące zapisem przebiegu symulacji poddano analizie skupień za pomocą programu HCPM (Gront & Kolinski 2005), następnie z trzech najliczniejszych klastrów wyselekcjonowano struktury reprezentatywne dla danego klastra, mierząc wartość RMSD wszystkich elementów klastra względem jego centroidu i wybiarając tę, dla której wartość RMSD była najniższa.

W trzech spośród 10 przypadków otrzymane modele charakteryzowały się dokładnością porównywalną ze strukturami krystalograficznymi (RMSD < 1 Å). W przypadku kolejnych dwóch, wartość RMSD względem struktury krystalograficznej wyniosła poniżej 2.5 Å. W pozostałych 5 przypadkach wartości RMSD dla końcowych modeli zawierały się w przedziale 5-10 Å. Przyczyną tych

nieco większych niedokładności było przyłączenie się cząsteczek kofaktorów we właściwym miejscu na powierzchni receptorów, lecz w niewłaściwej orientacji.



Rys. 4.1

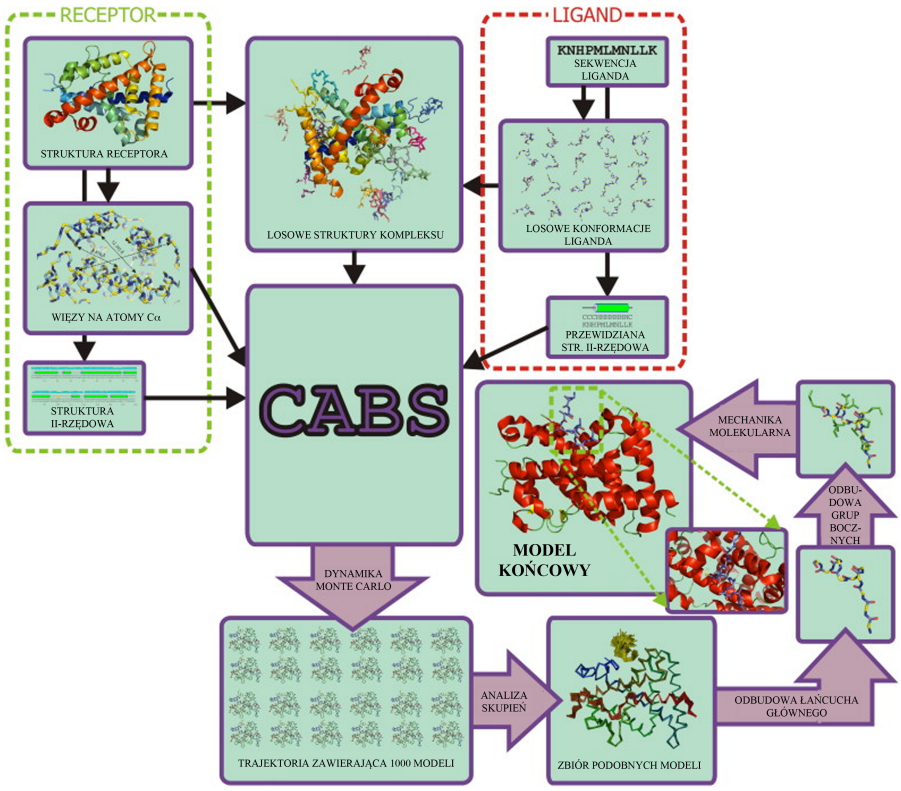
Przewidziane modele (gruba linia) nałożone na struktury krystalograficzne (cienka linia). Strzałki wskazują na cząsteczki kofaktorów A) kod pdb: 1RJK; rmsd względem struktury krystalograficznej 0.78 Å. B) kod pdb: 1NQ7; rmsd 0.47Å

W pracy P.I zaprezentowano możliwości algorytmu giętkiego dokowania opartego na modelu CABS. Wykazano, że dla krótkich peptydów możliwe jest bardzo efektywne przeszukiwanie przestrzeni konformacyjnej układu. W każdym z badanych przykładów algorytm był w stanie poprawnie zlokalizować miejsce wiązania na powierzchni receptora, bez żadnych danych naprowadzających, czym wyróżnia się spośród innych znanych metod dokowania. Jednocześnie otrzymane wyniki wskazały na tendencji algorytmu do generowania tzw. „false positives" – dokowania liganda we właściwym miejscu, lecz w niewłaściwej konformacji. Przyczyn tego zjawiska należy poszukiwać w potencjale kontaktowym grup bocznych, który pierwotnie został stworzony do modelowania oddziaływań wewnątrz-białkowych, a następnie został jedynie zaadaptowany w algorytmie dokowania. Wyprowadzenie specyficznego potencjału dla oddziaływań między białkowych powinno poprawić jakość otrzymywanych modeli.

1.2 Wieloskalowe modelowanie trójwymiarowych struktur kompleksów białkowych

Przewidywanie struktur kompleksów białkowych jest nie tylko interesującym zagadnieniem akademickim, ale posiada istotne praktyczne zastosowanie w procesie racjonalnego projektowania leków. Aby to było możliwe konieczne jest generowanie modeli wysokiej rozdzielczości, pozwalających na badanie mechanizmów jakimi kierują się białka w funkcjonowaniu organizmów.

Model CABS generuje modele w reprezentacji zredukowanej jedynie do atomów węgla Cα. Uproszczenie reprezentacji umożliwia modelowanie układów o znacznych rozmiarach oraz relatywnie długich procesów, które w tych układach zachodzą. Podobne symulacje z użyciem dynamiki molekularnej i mechaniki kwantowej nie są możliwe z powodu ogromnych kosztów obliczeniowych. Wysoka efektywność modelu CABS odbywa się kosztem rozdzielczości uzyskiwanych modeli. W reprezentacji ograniczonej jedynie do atomów Cα są one zbyt ogólne, aby stanowić wiarygodne źródło danych dla badań nad mechanizmami wielu reakcji.

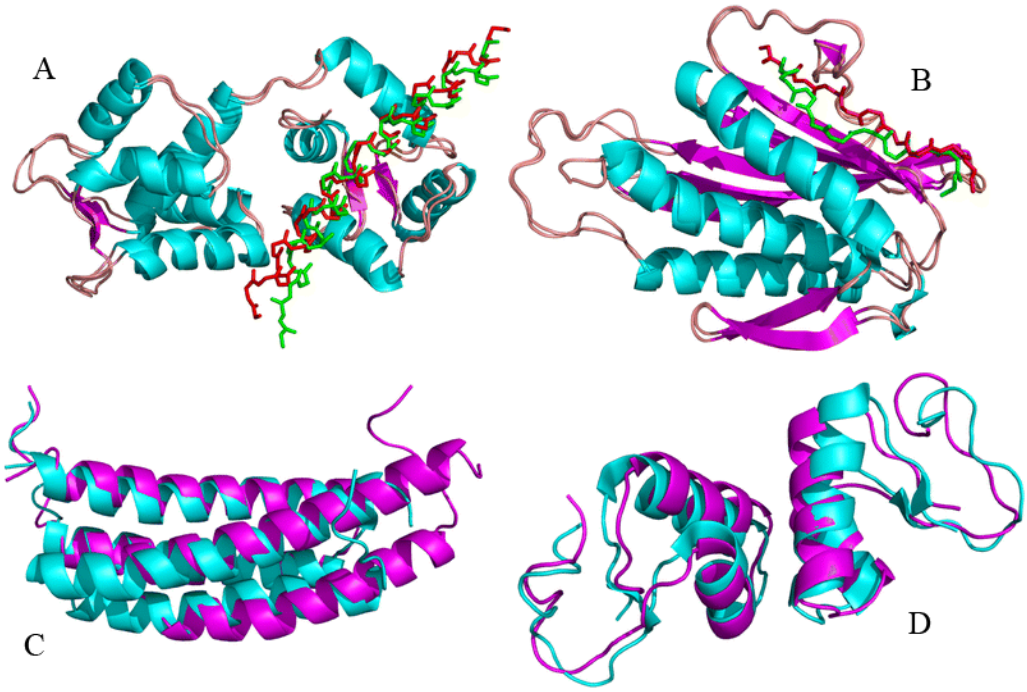


Rys 4.2

Schemat modelowania wielkoskalowego przy użyciu modelu CABS. Ramki pokazują przepływ danych, natomiast strzałki odpowiadają kolejnym procedurom.

W pracy P.II przedstawiono schemat modelowania wielkoskalowego, którego centralnym punktem jest symulacja dokowania za pomocą CABS (Rys 4.2). Procedura łączy w sobie wiele narzędzi, pozwalających generować modele w pełno-atomowej rozdzielczości. Działanie procedury przetestowano na jedenastu kompleksach białkowych, dla których znane są struktury krystalograficzne. Osiem przynależy do klasy białko-peptyd, a pozostałe trzy to niewielkie homodimery.

Schemat modelowania składa się z wielu etapów, które zostały szczegółowo opisane w pracy P.II. Jedyne dane wymagane do działania procedury to sekwencja aminokwasów w cząsteczkach receptora i liganda. Możliwe jest jednak wykorzystanie dodatkowych informacji (struktura receptora, przewidziana struktura II-rzędowa, więzy odległości), dzięki którym modelowanie jest dokładniejsze. W przedstawionych w pracy P.II przykładach wykorzystano struktury natywne receptorów i sekwencje ligandów.



Rys. 4.3

Rysunek przedstawia przewidziane struktury kompleksów nałożone na struktury krystalograficzne. A i B to przykłady modelowania białko-peptyd (modele w kolorze czerwonym; struktury krystalograficzne w kolorze zielonym), natomiast C i D to homodimery (modele w kolorze niebieskim, struktury krystalograficzne w kolorze fioletowym).

Otrzymane wyniki potwierdzają możliwości zarówno modelu CABS do uzyskania modeli wysokiej jakości, jak i całej procedury do skutecznego wyboru końcowych modeli. We wszystkich ośmiu przypadkach typu białko-peptyd uzyskano dokładność porównywalną z modelami krystalograficznymi. Dla trzech homodimerów zaobserwowano spadek dokładności modeli postępujący wraz ze wzrostem rozmiaru układu od RMSD równego 1.69 Å dla modelu zamka leucynowego[23] (62 aminokwasy), do 5.21 Å dla homodimeru białka ROP[24] (126 aminokwasów).

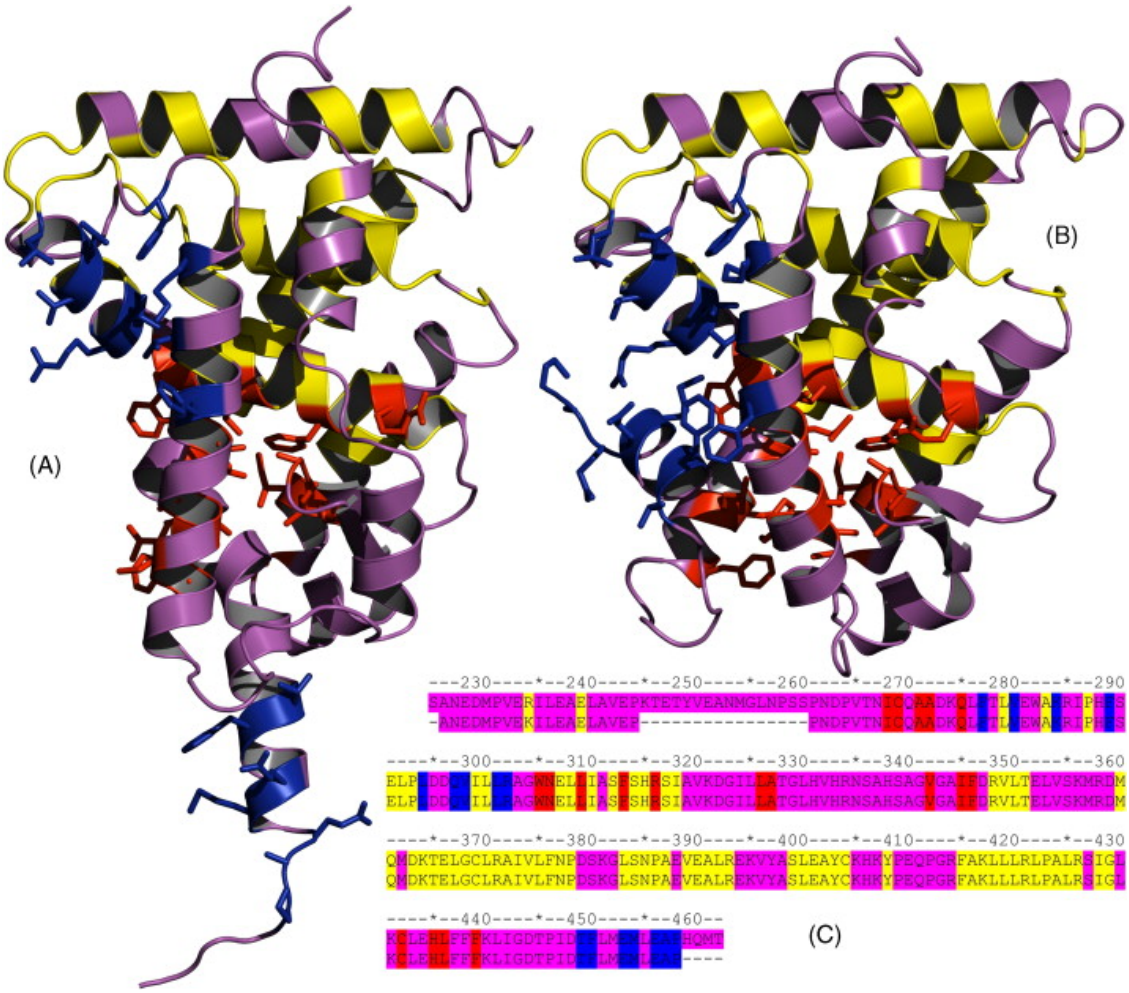
1.3 Badanie mechanizmu aktywacji receptora retinoidów α (RXRα) przez cząsteczki kwasu 9-cis retinowego i koaktywatora TRAP220

Receptor α (RXRα) kwasu 9-cis retinowego jest białkiem należącym do rodziny receptorów jądrowych. Białko to odgrywa ważną rolę w regulacji ekspresji genów. W formie niezwiązanej z agonistą RXRα łączy się z korepresorami, natomiast po przyłączeniu agonisty miejsce korepresora zajmuje w kompleksie koaktywator, co powoduje rozpoczęcie transkrypcji genu.

W pracy P.III zbadano mechanizm aktywacji RXRα przez cząsteczki kwasu 9-cis retinowego i peptydu będącego fragmentem koaktywatora TRAP220. Źródłem danych były dwie struktury krystalograficzne z bazy PDB: 1lbd (apo-RXRα) i 1xdk:AE (holo-RXRα związane z peptydem TRAP220). Wartość RMSD pomiędzy tymi strukturami wynosi ponad 8Å, co wskazuje na znaczną zmianę konformacji wynikającą z przyłączenia agonisty.

[23] Zamek leucynowy to często spotykany motyw strukturalny, występujący najczęściej w czynnikach transkrypcyjnych, w miejscu w którym łączą się one z nicią DNA. Zamek leucynowy składa się z dwóch równoległych helis, w których na co siódmej pozycji występuje aminokwas Leucyna. Ustawione naprzeciw siebie grupy boczne Leucyn oddziałują ze sobą hydrofobowo „zamykając zamek".

[24] ang. Repressor of Primer



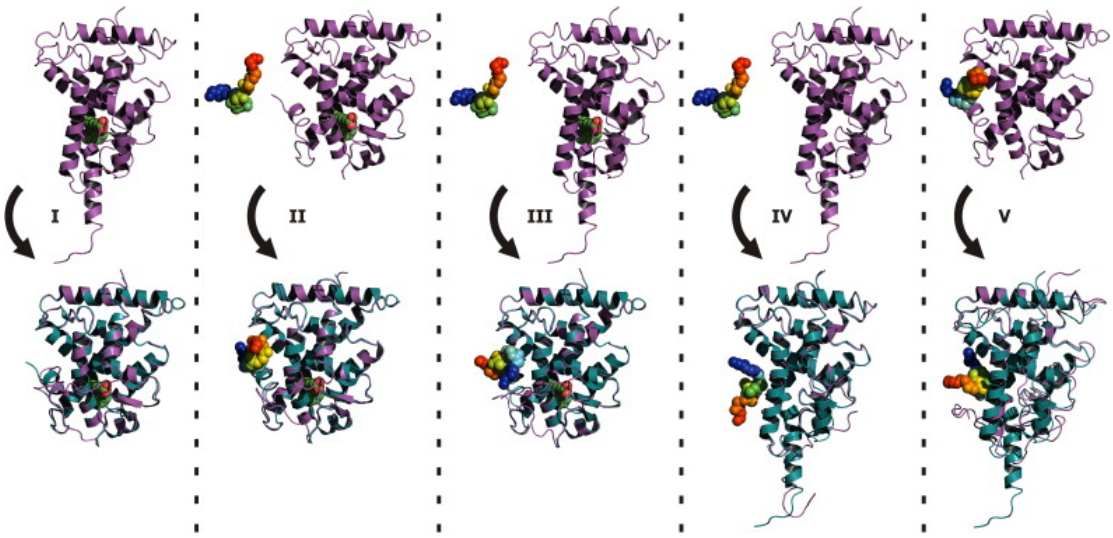
Rys. 4.4

Struktura krystalograficzna RXRα w formie apo (A) i holo (B). Reszty, które w formie apo kontaktują się z cząsteczką kwasu retinowego zaznaczono w kolorze czerwonym, te w kontakcie z koaktywatorem – w kolorze niebieskim. W kolorze żółtym zaznaczono reszty najsilniej zakonserwowane. (C) Uliniowienie sekwencji w kolorach jak wyżej.

Przeprowadzono 5 symulacji reprezentujących różne scenariusze, które mogą zachodzić w procesie aktywacji RXRα:

1) reorientacja receptora w obecności agonisty
2) przyłączenie kofaktora do kompleksu receptor/agonista
3) jednoczesna reorientacja receptora w obecności agonisty i przyłączenie kofaktora
4) przyłączenie kofaktora do receptora w formie apo
5) transformacja receptora holo ➔ apo po oddysocjowaniu agonisty

Celem pracy było zbadanie mechanizmu aktywacji RXRα, w szczególności kolejności w jakiej przyłączają się do niego agonista i koaktywator. Symulacje 1) i 2) odwzorowują kolejność „najpierw agonista, potem koaktywator", symulacja 3) – „agonista i koaktywator jednocześnie", symulacja 4) – pierwszy etap w sekwencji „najpierw koaktywator, potem agonista".



Rys. 4.5

Początkowe i końcowe konformacje modelowanych układów. W górnym rzędzie początkowe struktury receptora zaznaczono w kolorze fioletowym, kofaktor jako kolorowe sfery, a cząsteczkę kwasu 9-cis retinowego jako zielone sfery. W dolnym rzędzie na fioletowo zaznaczono końcowe modele receptora nałożone na odpowiednie struktury krystalograficzne.

Model CABS pozwala na symulowanie jedynie molekuł będących polipeptydami. Dlatego też obecność, bądź nieobecność agonisty została odwzorowana za pomocą szczególnego wykorzystania więzów odległości narzuconych na niektóre reszty receptora. W pierwszej kolejności na podstawie struktury receptora w formie holo zidentyfikowano które reszty znajdują się w bezpośrednim sąsiedztwie agonisty. Następnie zmierzono wzajemne odległości pomiędzy tymi resztami i narzucono silne więzy, co pozwalało zachować konformacje otoczenia agonisty, tak jakby się tam rzeczywiście znajdował.

Modele otrzymane w wyniku symulacji porównano ze strukturami krystalograficznymi. W przypadku symulacji 1 i 2 odtworzono symulowany proces z bardzo wysoką dokładnością, co wskazuje na sekwencję aktywacji RXRα „najpierw agonista, potem kofaktor". Jednocześnie wysokie wartość RMSD dla modeli

otrzymanych w wyniku symulacji 3 i 4 pokazują, że inna sekwencja zdarzeń nie zachodzi. W symulacji 5 modelowano proces dysocjacji koaktywatora i przejście holo-apo receptora po usunięciu agonisty. Procesu tego nie udało się odwzorować z zadowalającą dokładnością – cząsteczka kofaktora nie oderwała się od powierzchni receptora, który z kolei nie wrócił w pełni do formy apo. Przyczyny należy doszukiwać się w tendencji algorytmu CABS do generowania tzw. „false positives", wynikającej prawdopodobnie z braku specyficznego potencjału oddziaływań białko-białko.

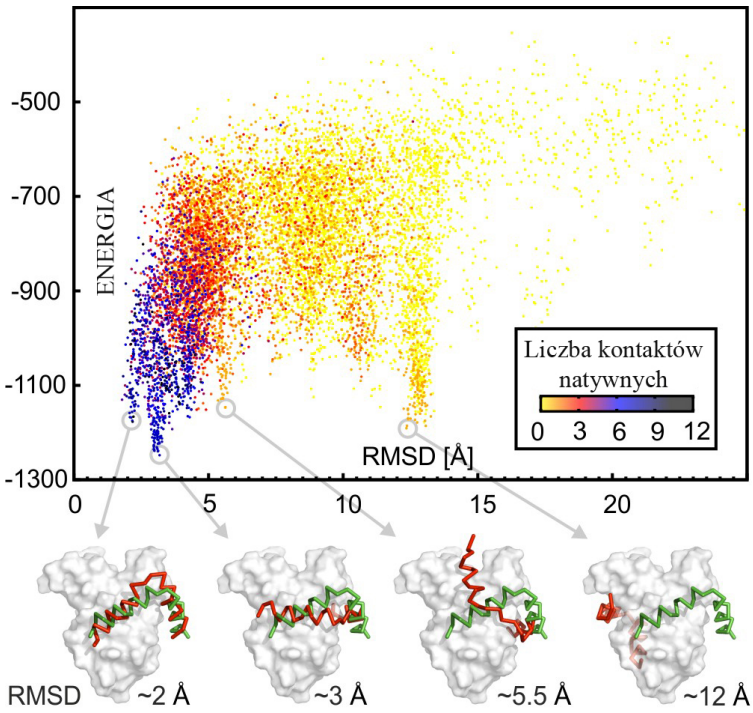
Zaproponowana metodologia pozwoliła na precyzyjne odtworzenie procesu aktywacji receptora retinoidów RXRα. Podobne podejście może być stosowane do badania innych układów i stanowi szybką i tanią alternatywę lub uzupełnienie dla badań eksperymentalnych.

1.4 Badanie mechanizmu jednoczesnego zwijania się białka nieustrukturyzowanego pKID w trakcie tworzenia kompleksu z domeną KIX białka CREB

Gdy występują w izolacji, białka nieustrukturyzowane charakteryzują się brakiem stabilnej struktury trzeciorzędowej w warunkach fizjologicznych. Brak jednoznacznej struktury może dotyczyć całej cząsteczki, bądź tylko pewnych jej fragmentów. Białka nieustrukturyzowane pełnią ważne funkcje w organizmach żywych, często regulują funkcjonowanie innych białek. Przyjmują one konkretną strukturę trzeciorzędową dopiero po związaniu się z partnerem.

Mechanizmy według których białka nieustrukturyzowane pełnią swoje funkcje są przedmiotem wielu badań zarówno eksperymentalnych (Parker et al. 1996; Radhakrishnan et al. 1997; Radhakrishnan et al. 1998; Shaywitz et al. 2000; Brüschweiler et al. 2013), jak i teoretycznych (Turjanski et al. 2008; Ganguly & Chen 2009; Espinoza-Fonseca 2009; Chen 2009; Huang & Liu 2010; Dadarlat & Skeel 2011; Ganguly & Chen 2011). Ostatnie badania wskazują na dwustopniowy proces, w wyniku którego białka nieustrukturyzowane łączą się z partnerem i przyjmują aktywną konformację. Najpierw tworzą „luźny" kompleks za pomocą słabych, niespecyficznych oddziaływań pozwalających na rozpoznanie się molekuł na dużych odległościach. Następnie przeorganizowują swoją strukturę do konformacji aktywnej

z jednoczesnym wytworzeniem natywnych kontaktów z partnerem, przy czym drugi etap jest zdecydowanie wolniejszy od pierwszego. Przypomina to mechanizm nukleacji-kondensacji, zgodnie z którym zwija się wiele jednodomenowych białek globularnych (Itzhaki et al. 1995; Fersht 1997).



Rys. 4.6

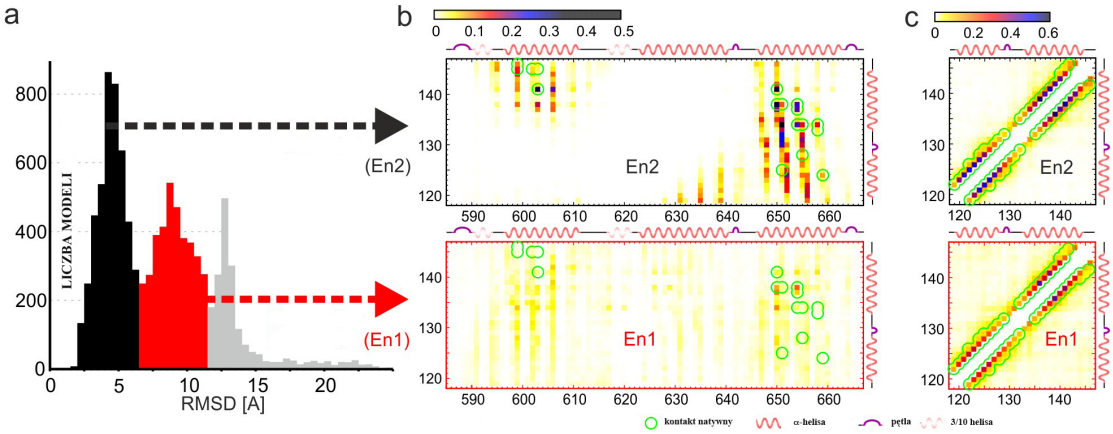
Zbiorcza charakterystyka 10 trajektorii, które zbiegły do blisko-natywnych konformacji. Energia wyrażona jest w wewnętrznych jednostkach modelu CABS. Pod wykresem pokazano przykładowe struktury o najniższych energiach z niektórych minimów (na czerwono oznaczono model, na zielono strukturę natywną).

Kompleks pKID-KIX jest wnikliwie zbadanym modelowym przykładem, w którym białko nieustrukturyzowane pKID zwija się do aktywnej konformacji w trakcie łączenia się z domeną KIX. pKID składa się z dwóch helis połączonych pętlą, które w formie aktywnej ustawiają się prawie prostopadle do siebie.

W pracy P.IV przeprowadzono szereg symulacji, w których cząsteczka pKID startuje z losowej konformacji i losowego miejsca w pobliżu domeny KIX. 10 trajektorii, które zbiegły do blisko-natywnych konformacji zostało wspólnie przedstawionych na rysunku (Rys. 4.6). Liczne minima wskazują na nietrwały charakter powstających kompleksów.

Dystrybucja wartości RMSD pokazana na rysunku (Rys. 4.7) pozwoliła na podzielenie wygenerowanych modeli na trzy klasy odpowiadające kolejnym etapom formowania się kompleksu:

- RMSD > 11.5 Å – stan niezwiązany bądź kompleks niewłaściwy (pKID związane w złym miejscu na powierzchni domeny KIX), 19% modeli;
- 6.5 Å< RMSD < 11.5 Å – kompleks typu En1, niewłaściwy bądź częściowo właściwy (pKID związane w sąsiedztwie poprawnego miejsca wiązania na powierzchni domeny KIX), 37% modeli;
- RMSD < 6.5 Å – kompleks typu En2, właściwy bądź częściowo właściwy (pKID związane przynajmniej częściowo w poprawnym miejscu na powierzchni receptora), 44% modeli.



Rys 4.7

Charakterystyka dominujących elementów w strukturach kompleksów. (a) Dystrybucja wartości RMSD wskazuje na istnienie trzech zbiorów kompleksów. (b) Mapa częstości kontaktów dla pKID i KIX. (c) Mapa kontaktów wewnętrznych pKID. Kontakty zliczono używając promienia obcięcia wynoszącego 5 Å pomiędzy środkami ciężkości grup bocznych. W (c) kontakty sąsiadów w sekwencji (i+1 i i+2) zostały pominięte dla przejrzystości.

Analiza otrzymanych map kontaktów pozwoliła sformułować mechanizm tworzenia kompleksu pKID-KIX, w którym kluczowe role na etapie utworzenia luźnego kompleksu odgrywają hydrofobowe reszty pKID: Ile137, Leu138 i Leu141, co wydaje się być czynnikiem charakterystycznym wyłącznie dla białek nieustrukturyzowanych. W typowych układach receptor – ligand rozpoznanie się molekuł zachodzi za pomocą oddziaływań pomiędzy naładowanymi resztami, gdyż oddziaływania elektrostatyczne rozciągają się na dalsze odległości niż oddziaływania

hydrofobowe. Tym samym mechanizm tworzenia kompleksu pKID-KIX przypomina mechanizm nukleacji-kondensacji, charakterystyczny dla zwijania się pojedynczych domen białek.

1.5 Modelowanie zjawiska mimikry molekularnej gangliozydu GD2 przez grupę peptydów w kompleksie z przeciwciałem 14G2a

Nerwiak płodowy to groźny nowotwór złośliwy atakujący najczęściej dzieci w wieku niemowlęcym. Mimo iż relatywnie łatwo jest go zdiagnozować, przeżywalność pozostaje na poziomie około 50%. Z występowaniem nerwiaka płodowego stowarzyszona jest nadekspresja gangliozydu GD2, co stwarza możliwość stosowania immunoterapii. Jednocześnie immunogenność GD2 jest znikoma (Bolesta et al. 2005), zwłaszcza u niemowląt, przez co podjęto próbę opracowania szczepionki opartej na mimotopach GD2. W tym celu przeprowadzono skrining biblioteki peptydów składających się z 12 aminokwasów pod kątem zdolności do przyłączania się do przeciwciała 14G2a (Horwacik et al. 2007). W wyniku eksperymentu zidentyfikowano 5 peptydów o porównywalnym do GD2 powinowactwie do 14G2a oraz wykazano, że warunkiem koniecznym do ich związania przez przeciwciało jest występowanie w peptydzie na pozycjach 2 i 11 aminokwasów Cystein połączonych mostkiem siarczkowym. W tabeli 3.1 przedstawiono sekwencje wybranych peptydów.

W pracy P.V przeprowadzono analizę 5 wyselekcjonowanych wcześniej peptydów w celu zbadania ich aktywności w zależności od struktury. W tym celu przeprowadzono analizę cytometryczną oraz kompetencyjny test immunoenzymatyczny na badanych peptydach. Testowano ich powinowactwo nie tylko do przeciwciała 14G2a, ale również względem innych specyficznych dla gangliozydów przeciwciał. Badania wykazały, że peptydy #65, #94 i #85 współdzielą to samo miejsce aktywne na powierzchni przeciwciała, mimo istotnych różnic w ich sekwencjach. Jednocześnie potwierdziła się kluczowa dla wiązania, rola mostka siarczkowego pomiędzy resztami 2 i 11 peptydów. Ponadto okazało się, że peptydy #65, #94 i #85 nie wchodzą w reakcję krzyżową z innymi niż 14G2a specyficznymi dla gangliozydów przeciwciałami.

Badania *in vitro* zostały również wsparte modelem teoretycznym oddziaływania peptydów z przeciwciałem. W pierwszej kolejności opracowany został

model przeciwciała 14G2a, gdyż jego struktura nie była znana. Przeprowadzono modelowanie z wykorzystaniem szablonu, który został zidentyfikowany i uliniowiony z sekwencją 14G2a za pomocą programu PSI-BLAST. Model został skonstruowany za pomocą programu CABS. Następnie, również z wykorzystaniem modelu CABS, przeprowadzono giętkie dokowanie badanych peptydów do otrzymanego modelu. Na cząsteczkę przeciwciała nałożono więzy pochodzące z szablonu, utrzymujące receptor w konformacji bliskiej natywnej, natomiast na cząsteczkach dokowanych peptydów wymuszono istnienie mostka siarczkowego pomiędzy cysteinami. Trajektorie otrzymane z symulacji zostały poddane analizie skupień, a następnie struktury reprezentatywne z największych klastrów odbudowano do pełnoatomowej reprezentacji. W następnym kroku przeprowadzono minimalizację energii w programie Sybyl, używając pola siłowego AMBER99. W poniższej tabeli zamieszczono wartości energii wiązania peptydów.

Kod peptydu	Sekwencja	Energia wiązania
#8	NCDLLTGPMLCV	-170.5 kcal/mol
#85	VCNPLTGALLCS	-165.1 kcal/mol
#D	GCDALSGHLLCS	-126.7 kcal/mol
#65	SCQSTRMDPNCW	-228.9 kcal/mol
#94	RCNPNMEPPRCF	-178.9 kcal/mol

Tabela 4.1

W tabeli zamieszczono sekwencje aminokwasowe peptydów imitujących gangliozyd GD2. Wszystkie składają się z 12 reszt aminokwasowych oraz zawierają aminokwas Cysteinę na pozycjach 2 i 11. Dodatkowo w tabeli zamieszczono wartości wyznaczonej energii wiązania peptydów z przeciwciałem.

Wnioski z modelowania potwierdzają wyniki badań eksperymentalnych. We wszystkich 5 przypadkach badane peptydy zostały zadokowane w tym samym miejscu na powierzchni receptora, mimo iż dane wejściowe nie zawierały żadnych informacji o miejscu aktywnym. Wartości energii wiązania wskazują, że najsilniej do przeciwciała wiąże się peptyd #65, co również zgadza się jakościowo z badaniami eksperymentalnymi. Analogiczne symulacje dokowania przeprowadzono również dla peptydów bez narzuconego warunku o istnieniu mostka siarczkowego. Analiza skupień przeprowadzona na otrzymanych w ten sposób symulacjach wykazała

znaczną ilość niewielkich klastrów o niskiej gęstości, przez co niemożliwe było jednoznaczne wybranie pojedynczej struktury reprezentującej wynik dokowania. Pozwala to wysnuć wniosek o kluczowej roli mostka siarczkowego dla mimikry gangliozydu GD2 przez badane peptydy.

2 Dokowanie białko-białko

2.1 Teoretyczny model kompleksu białkowego ludzkiej Telomerazy

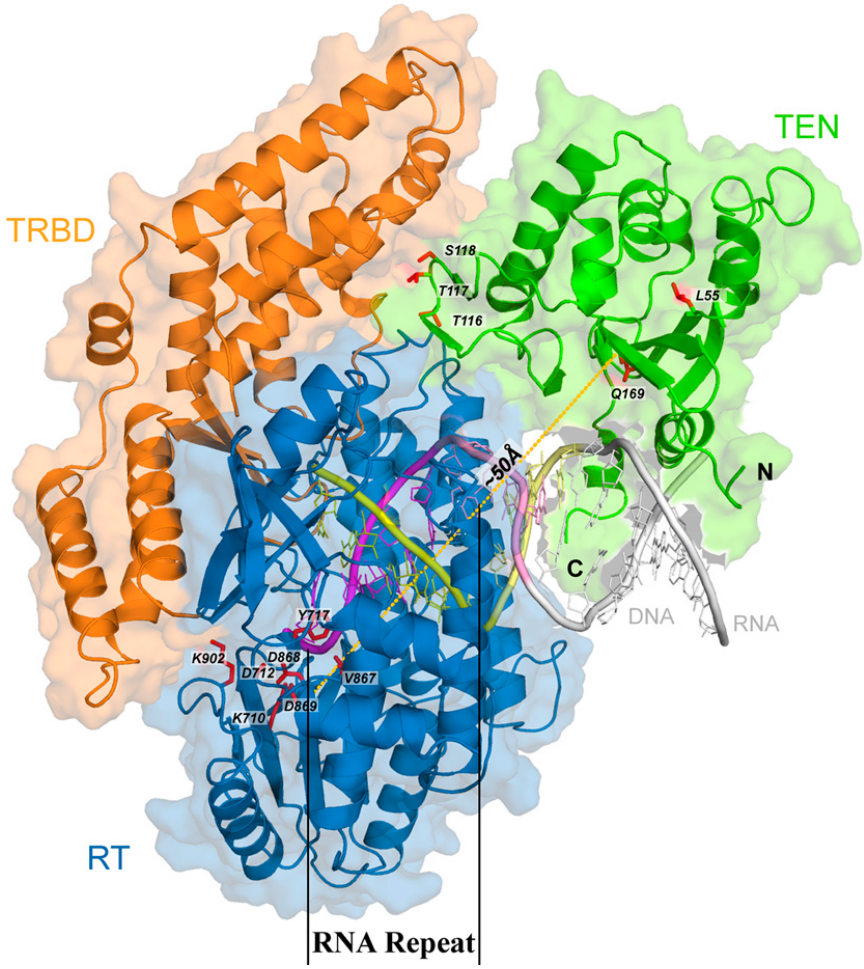
Podczas replikacji DNA wstęga kwasu nukleinowego przesuwana jest tylko w jednym kierunku – polimeraza DNA może przemieszczać się jedynie od końca 3' w kierunku końca 5' budując nową nić. Z tego powodu w komórkach eukariotycznych jedna z nici replikowanego DNA (nić wiodąca) kopiowana jest w sposób ciągły, a druga (nić opóźniona) skokowo poprzez przyłączanie tzw. fragmentów Okazaki do startera replikacji. Na samym końcu 3' nici opóźnionej występuje tzw. problem końca replikacji – fragment DNA nie jest kopiowany.

Natura znalazła sposób na poradzenie sobie z tym problemem. Telomeraza to enzym rybonukleinowy, który przyłącza krótkie fragmenty niekodującego DNA (telomery) to 3' końca kopiowanej nici. Każda kolejna replikacja skraca obszar telomerów, lecz tym samym kodujące fragmenty DNA pozostają niezagrożone. Jednocześnie jest to mechanizm kontrolujący wiek komórki. Z czasem telomerów zaczyna brakować, co prowadzi do błędów w trakcie replikacji, a w efekcie do śmierci komórki. Możliwość kontrolowania czasu życia komórki sprawia, że telomeraza jest doskonałym celem dla leków np. antynowotworowych.

Telomeraza została odkryta w 1985 roku przez Carol Greider i Elizabeth Blackburn (Greider & Blackburn 1985), za co otrzymały w 2009 roku Nagrodę Nobla, lecz pomimo usilnych prac nad uzyskaniem trójwymiarowej struktury, do dziś nie udało się jej uzyskać. W pracy P.VI zaprezentowano pierwszy teoretyczny model kompleksu ludzkiej telomerazy wraz z obszarem telomeru pojedynczej nici DNA.

Kompleks telomerazy składa się z fragmentu RNA kodującego sekwencje telomerów hTR (**h**uman **T**elomerase **R**NA) oraz z białka TERT (**T**elomerase **R**everse **T**ranscriptase). TERT składa się 1132 reszt aminokwasowych, zorganizowanych w 4 funkcjonalne domeny: TEN (**T**elomerase **E**ssential **N**-terminal Domain), TRBD (**T**elomerase **R**NA **B**inding **D**omain), RT (**R**everse **T**ranscriptase domain) i CTE (**C**-

Terminal **E**xtension). W bazie PDB znajdują się krystalograficzne struktury domen TRBD i RT chrząszcza Trojszyka gryzącego (*Tribolium castaneum*) w kompleksie z RNA i DNA (Gillis et al. 2008; Mitchell et al. 2010) oraz domeny TEN (Jacobs et al. 2006) i TRBD (Rouda & Skordalakes 2007) z organizmu pierwotniaka *Tetrahymena thermophilla*.



Rys. 4.8

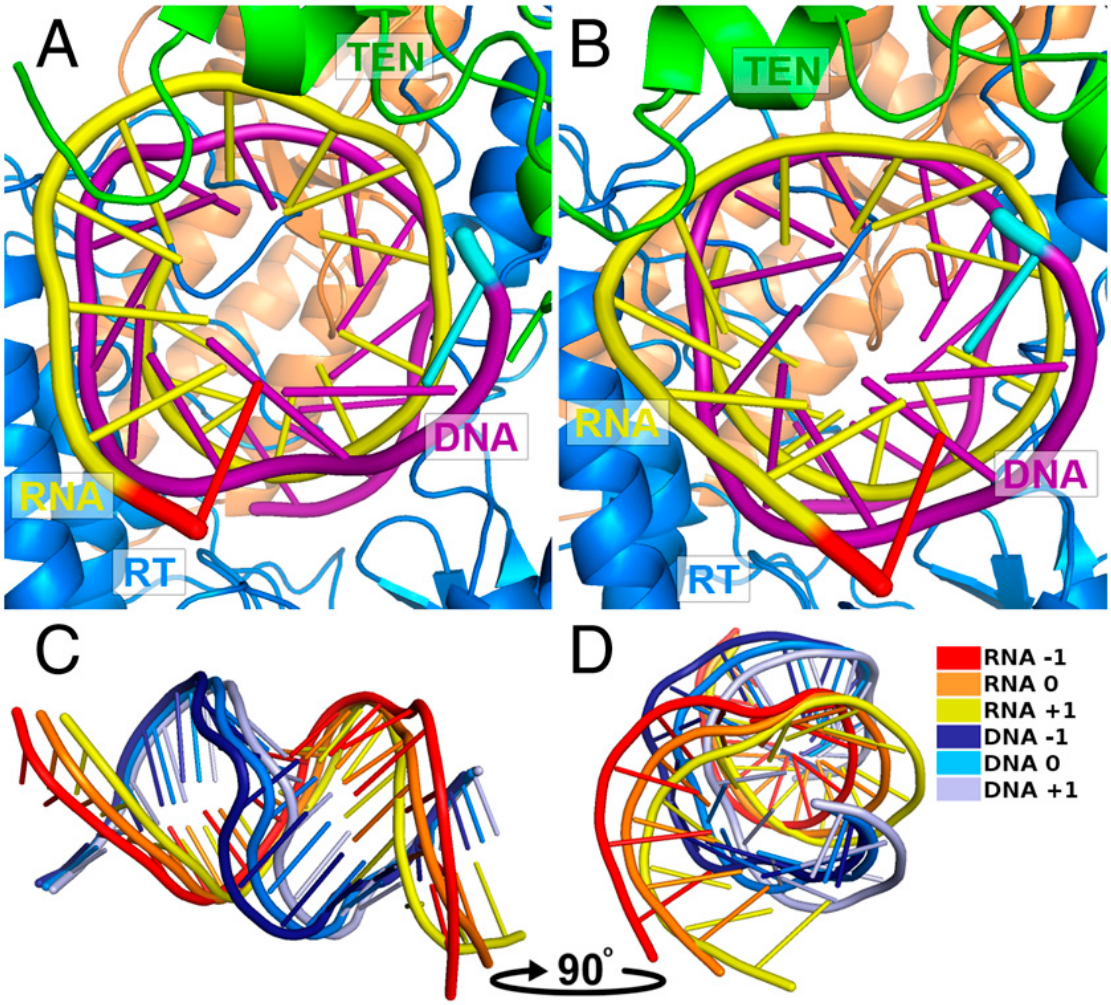
Teoretyczny model kompleksu telomerazy. W kolorze żółtym zaznaczono cząsteczkę RNA stanowiącą szablon, według którego do nici DNA (kolor różowy) przyłączane są kolejne nukleotydy.

Podobieństwo sekwencyjne pomiędzy dostępnymi strukturami, a odpowiednimi fragmentami sekwencji ludzkiej telomerazy wynosi 20-25%, przez co niemożliwe było zastosowanie standardowych metod modelowania porównawczego. Wykorzystano zaawansowaną metodę opartą na badaniu profili sekwencyjnych Meta-Basic (Ginalski et al. 2004) oraz program Modeller (Eswar et al. 2007) do skonstruowania w oparciu o homologię modeli domen TEN, TRBD oraz

RT. Następnie za pomocą procedury CABSDock skonstruowano model kompleksu 3 domen. W tym celu przeprowadzono trzy symulacje dokowania dla każdej możliwej pary domen z trójki TEN, TRBD i RT. Otrzymano wiele alternatywnych modeli obejmujących po dwie domeny Telomerazy. Następnie modele zostały na siebie nałożone, tak aby te same domeny w różnych kompleksach (np. domena TEN w kompleksie TEN/TERT i w kompleksie TEN/RT) pokrywały się. Z uzyskanego w ten sposób zbioru odrzucono modele, w których wystąpiły zawady steryczne. Pozostałe modele zostały poddane jeszcze raz symulacji w programie CABSDock w celu zrelaksowania układu. W wyniku analizy skupień uzyskanej trajektorii zaproponowano najbardziej prawdopodobny model trzech domen telomerazy. Ze struktury domeny TRBD telomerazy z organizmu chrząszcza *Tribolium Castaneum* przekopiowano fragmenty nici szablonowego RNA oraz telomeru DNA. Tak uzyskany kompletny model został poddany minimalizacji energii za pomocą programu SYBYL z użyciem pola siłowego AMBER.

Na podstawie sporządzonego modelu telomerazy zaproponowano mechanizm działania kompleksu. W tym celu przeprowadzono symulacje z użyciem modelu ANM[25] (Atilgan et al. 2001; Yang et al. 2007), w którym spośród wszystkich atomów Cα, P i O4' połączono pary znajdujące się od siebie nie dalej niż 13Å, za pomocą identycznych sprężyn. Analiza drgań normalnych układu pozwoliła sformułować mechanizm działania telomerazy, w którym po ukończeniu cyklu syntezy pojedynczego fragmentu telomeru, szablonowe RNA jest przesuwane do kolejnej zasady w łańcuchu DNA. Następnie fragment domeny TER, wiążący RNA przesuwa się względem nici DNA, utrzymując heteroduplex RNA:DNA. Nowo zsyntezowany fragment DNA jest uzupełniany przez komplementarny fragment nici, podczas gdy szablonowe RNA zostaje wyeksponowane do rozpoczęcia kolejnego cyklu syntezy. Na Rysunku 4.9 przedstawiono strukturalny model ruchów kompleksu.

[25] ANM – Anisotropic Network Model



Rys. 4.9

Strukturalny model ruchów ludzkiej telomerazy, uzyskany z analizy drgań układu w modelu ANM. Ewolucja układu w kierunku dodatnim (A) i ujemnym (B) głównego modu drgań wskazuje na rotacje i przesunięcie heteroduplexu RNA:DNA. (C) i (D) ukazują trzy stany heteroduplexu: wyjściowy stan (0) oraz deformacje w kierunkach ujemnym (-1) i dodatnim (+1) głównego modu drgań.

V Podsumowanie i wnioski końcowe

Celem pracy doktorskiej było opracowanie opartej na modelu CABS metody służącej do przewidywania struktur kompleksów białkowych oraz do modelowania ich dynamiki. Unikalną cechą opracowywanej metody miała być możliwość przeprowadzenia symulacji dokowania przy zachowaniu pełnej giętkości molekuł. Jednocześnie pragnieniem autora było zaprojektowanie narzędzia uniwersalnego, stosowalnego zarówno do układów białko-peptyd, jak i do wielodomenowych kompleksów białkowych.

W pracy przedstawiono szczegółowy opis konstrukcji automatycznej procedury CABSDock oraz zaprezentowano przykłady jej praktycznego zastosowania. Dzięki uniwersalności i wysokiej efektywności zastosowanie CABSDock otwiera nowe możliwości badania oddziaływań pomiędzy białkami, na skalę niedostępną dzisiaj dla badań eksperymentalnych.

Opis procedury zawiera dokładną charakterystykę poszczególnych narzędzi wchodzących w skład CABSDock. W przypadku dokowania białko-białko, pierwszym krokiem jest przybliżone dokowanie za pomocą programu FTDOCK. Składniki kompleksu, traktowane są jako bryły sztywne, a ich dopasowanie badane jest jedynie na podstawie komplementarności kształtów ich powierzchni. W kolejnym kroku wstępne modele są oceniane pod kątem oddziaływań pomiędzy molekułami. Najlepiej ocenione modele są następnie poddawane symulacji dynamiki w modelu CABS. Ostatni krok to wybór najbardziej prawdopodobnych modeli oraz rekonstrukcja ich detali strukturalnych.

Konstrukcja procedury CABSDock wymagała od autora rozszerzenia modelu CABS, aby umożliwić symulacje wielu łańcuchów białkowych, przy jednoczesnym zachowaniu wysokiej efektywności oryginalnego programu. Ponadto opracowano zestaw narzędzi (skrypty do konwersji formatów plików wymaganych przez poszczególne programy oraz do analizy statystycznej otrzymywanych wyników) scalających poszczególne elementy CABSDock w automatyczną procedurę.

W celu przetestowania metody przeprowadzono eksperyment przewidywania struktur na grupie białek z rodziny receptorów jądrowych skompleksowanych z peptydami będącymi fragmentami czynników transkrypcyjnych. W 50%

przypadków otrzymano modele dobrej lub bardzo dobrej jakości. Jednocześnie wykazano zdolność algorytmu do bardzo efektywnego przeszukiwania powierzchni receptora pod kątem miejsca aktywnego dla liganda. We wszystkich przykładach peptydy zadokowane zostały we właściwym miejscy na powierzchni receptora jądrowego.

Zastosowano schemat wielkoskalowego modelowania struktury kompleksów białkowych obejmujących zarówno układy białko-peptyd, jak i kompleksy małych homodimerów. Badano peptydy większych rozmiarów (do 31 aminokwasów), niż w przypadku receptorów jądrowych oraz posiadające elementy regularnej struktury II-rzędowej (α-helisy i β-wstęgi).

Model CABS został również wykorzystany do badania mechanizmu aktywacji receptora retinoidów RXRα przez cząsteczkę kwasu retinowego oraz fragment koaktywatora. Zbadano różne scenariusze, które mogą zachodzić w trakcie tego procesu. Na podstawie otrzymanych wyników zaproponowano możliwy mechanizm aktywacji. Jednocześnie zaprezentowano metodę, dzięki której można szybko i tanio badać podobne procesy molekularne.

Zbadano również mechanizm jednoczesnego zwijania się białka nieustrukturyzowanego pKID w trakcie tworzenia kompleksu z jedną z domen białka CREB. Przeprowadzono szczegółową analizę ścieżki zwijania i stanów pośrednich takiego kompleksu. W konsekwencji zaobserwowano podobieństwo mechanizmu jednoczesnego zwijania i łączenia się białek do mechanizmu nukleacji-kondensacji, typowego dla zwijania się jednodomenowych białek globularnych.

Symulacje z wykorzystaniem modelu CABS posłużyły również jako wsparcie teoretyczne badań eksperymentalnych ukierunkowanych na poszukiwanie szczepionki przeciw nerwiakowi zarodkowemu (neuroblastoma). Badano stabilność kompleksów przeciwciała 14G2a z grupą peptydów imitujących obecność gangliozydu GD2, którego nadekspresja jest znakiem rozpoznawczym komórek zainfekowanych neuroblastoma.

Symulacje z użyciem CABSDock wykorzystano również do skonstruowania pierwszego teoretycznego modelu ludzkiej telomerazy – kompleksu białkowego składającego się z prawie 1200 reszt aminokwasowych, odpowiedzialnego *in vivo* za naprawianie błędów powstających w trakcie replikacji DNA. Przeprowadzono wielostopniowy eksperyment dokowania kolejnych domen telomerazy oraz

fragmentów nici RNA i DNA. Na podstawie uzyskanego modelu zaproponowano po raz pierwszy mechanizm działania tego kompleksu.

Podsumowując, stworzenie skutecznego i jednocześnie uniwersalnego narzędzia służącego do modelowania oddziaływań białek jest zadaniem bardzo trudnym. Ogromne rozmiary kompleksów białkowych i różnorodność ich potencjalnych oddziaływań sprawiają, że nie istnieje jedna prosta recepta mówiąca jak ten problem rozwiązać. W tej pracy zaproponowano podejście skoncentrowane przede wszystkim na wysokiej efektywności tak, aby z czasem mogło zostać zaimplementowane w postaci automatycznego serwera i wykorzystane do badania oddziaływań białek na skalę całych genomów. Dalsze prace będą ukierunkowane na wzbogacenie modelu CABS o możliwość badania odziaływań nie tylko między białkami, ale również innymi biomolekułami, w szczególności z kwasami nukleinowymi. Mimo iż problem giętkiego dokowania białek wciąż jest daleki od rozwiązania, na dzień dzisiejszy CABSDock wydaje się być bardzo dobrym narzędziem do automatycznego modelowania oddziaływań typu białko-peptyd oraz białko-małe białko, a w przypadku większych układów wciąż może być źródłem cennych i łatwych do uzyskania informacji.

VI Prace cytowane

Alberts, I.L., Todorov, N.P. & Dean, P.M., **2005**. Receptor flexibility in de novo ligand design and docking. *Journal of medicinal chemistry*, 48(21), 6585–96.

Alder, B.J. & Wainwright, T.E., **1959**. Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics*, 31(2), 459.

Alonso, H., Bliznyuk, A.A. & Gready, J.E., **2006**. Combining docking and molecular dynamic simulations in drug design. *Medicinal research reviews*, 26(5), 531–68.

Anfinsen, C.B., **1973**. Principles that govern the folding of protein chains. *Science*, 181, 223–230.

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O. & Bahar, I., **2001**. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1), 505–15.

Bae, W., Choi, M.-G., Hyeon, C., Shin, Y.-K. & Yoon, T.-Y., **2013**. Real-time observation of multiple-protein complex formation with single-molecule FRET. *Journal of the American Chemical Society*, 135(28), 10254–7.

Bahar, I., Chennubhotla, C. & Tobi, D., **2007**. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Current opinion in structural biology*, 17(6), 633–40.

Bakan, A. & Bahar, I., **2009**. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34), 14349–54.

Barril, X. & Fradera, X., **2006**. Incorporating protein flexibility into docking and structure-based drug design. *Expert Opinion on Drug Discovery*, 1(4), 335–349.

Bax, A. & Grzesiek, S., **1993**. Methodological advances in protein NMR. *Accounts of Chemical Research*, 26(4), 131–138.

Berman, H.M., Westbrook, J.D., Gabanyi, M.J., Tao, W., Shah, R., Kouranov, A., Schwede, T., Arnold, K., Kiefer, F., Bordoli, L., Kopp, J., Podvinec, M., Adams, P.D., Carter, L.G., Minor, W., Nair, R. & La Baer, J., **2009**. The protein structure initiative structural genomics knowledgebase. *Nucleic acids research*, 37(1), 365–8.

Bloch, F., **1946**. Nuclear Induction. *Physical Review*, 70(7-8), 460–474.

Bloembergen, N., Purcell, E. & Pound, R., **1947**. Nuclear magnetic relaxation. *Nature*, 160, 475–476.

Bloembergen, N., Purcell, E. & Pound, R., **1948**. Relaxation effects in nuclear magnetic resonance absorption. *Physical Review*, 73(7), 679–712.

Boehr, D.D., Nussinov, R. & Wright, P.E., **2009**. The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology*, 5(11), 789–96.

Bolesta, E., Kowalczyk, A., Wierzbicki, A., Rotkiewicz, P., Bambach, B., Tsao, C.-Y., Horwacik, I., Kolinski, A., Rokita, H., Brecher, M., Wang, X., Ferrone, S. & Kozbor, D., **2005**. DNA vaccine expressing the mimotope of GD2 ganglioside induces protective GD2 cross-reactive antibody responses. *Cancer research*, 65(8), 3410–18.

Bowie, J., Luthy, R. & Eisenberg, D., **1991**. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016), 164–170.

Brink, C., Hodgkin, D.C., Lindsey, J., Pickworth, J., Robertson, J.H. & White, J.G., **1954**. Structure of Vitamin B12: X-ray Crystallographic Evidence on the Structure of Vitamin B12. *Nature*, 174(4443), 1169–1171.

Brüschweiler, S., Konrat, R. & Tollinger, M., **2013**. Allosteric Communication in the KIX Domain Proceeds through Dynamic Repacking of the Hydrophobic Core. *ACS chemical biology*, 8(7), 1600–10.

Bujnicki, J.M., Elofsson, A., Fischer, D. & Rychlewski, L., **2001**. Structure prediction meta server. *Bioinformatics*, 17(8), 750–751.

Carlson, H.A., **2002**. Protein flexibility and drug design: how to hit a moving target. *Current Opinion in Chemical Biology*, 6(4), 447–452.

Carpenter, E.P., Beis, K., Cameron, A.D. & Iwata, S., **2008**. Overcoming the challenges of membrane protein crystallography. *Current opinion in structural biology*, 18(5), 581–6.

Cavasotto, C. & Abagyan, R., **2004**. Protein flexibility in ligand docking and virtual screening to protein kinases. *Journal of molecular biology*, 337(1), 209–225.

Chen, H. & Skolnick, J., **2008**. M-TASSER: an algorithm for protein quaternary structure prediction. *Biophysical journal*, 94(3), 918–28.

Chen, H.-F., **2009**. Molecular dynamics simulation of phosphorylated KID post-translational modification. *PloS one*, 4(8), 6516.

Chen, R., Mintseris, J., Janin, J. & Weng, Z., **2003**. A protein-protein docking benchmark. *Proteins*, 52(1), 88–91.

Chothia, C. & Lesk, A.M., **1986**. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4), 823–6.

Chou, J.J. & Sounier, R., **2013**. Solution nuclear magnetic resonance spectroscopy. *Methods in molecular biology*, 955, 495–517.

Claußen, H., Buning, C., Rarey, M. & Lengauer, T., **2001**. FlexE: efficient molecular docking considering protein structure variations. *Journal of molecular biology*, 308(2), 377–395.

Cox, M.J. & Weber, P.C., **1987**. Experiments with automated protein crystallization. *Journal of Applied Crystallography*, 20(5), 366–373.

Cozzini, P. & Kellogg, G., **2008**. Target Flexibility: An Emerging Consideration in Drug Discovery and Design†. *Journal of medicinal chemistry*, 51(20), 6237–6255.

Dadarlat, V.M. & Skeel, R.D., **2011**. Dual role of protein phosphorylation in DNA activator/coactivator binding. *Biophysical Journal*, 100(2), 469–77.

Das, R., André, I., Shen, Y., Wu, Y., Lemak, A., Bansal, S., Arrowsmith, C.H., Szyperski, T. & Baker, D., **2009**. Simultaneous prediction of protein folding and docking at high resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), 18978–83.

Davis, F.P. & Sali, A., **2005**. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, 21(9), 1901–7.

Dobbins, S.E., Lesk, V.I. & Sternberg, M.J.E., **2008**. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proceedings of the National Academy of Sciences of the United States of America*, 105(30), 10390–5.

Dominguez, C., Boelens, R. & Bonvin, A.M.J.J., **2003**. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7), 1731–7.

Dunbrack, R.L., **2006**. Sequence comparison and protein structure prediction. *Current Opinion in Structural Biology*, 16(3), 374–384.

Espinoza-Fonseca, L.M., **2009**. Thermodynamic aspects of coupled binding and folding of an intrinsically disordered protein: a computational alanine scanning study. *Biochemistry*, 48(48), 11332–4.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.-Y., Pieper, U. & Sali, A., **2007**. Comparative protein structure modeling using MODELLER. *Current protocols in protein science*, 2(9), 1–31.

Fernández-Recio, J. & Sternberg, M.J.E., **2010**. The 4th meeting on the Critical Assessment of Predicted Interaction (CAPRI) held at the Mare Nostrum, Barcelona. *Proteins: Structure, Function, and Bioinformatics*, 78(15), 3065–3066.

Fersht, A.R., **1997**. Nucleation mechanisms in protein folding. *Current Opinion in Structural Biology*, 7(1), 3–9.

Fischer, E., **1894**. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3), 2985–2993.

Fogolari, F., Brigo, A. & Molinari, H., **2002**. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *Journal of molecular recognition : JMR*, 15(6), 377–92.

Gabb, H.A., Jackson, R.M. & Sternberg, M.J., **1997**. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology*, 272(1), 106–20.

Ganguly, D. & Chen, J., **2009**. Atomistic details of the disordered states of KID and pKID. Implications in coupled binding and folding. *Journal of the American Chemical Society*, 131(14), 5214–23.

Ganguly, D. & Chen, J., **2011**. Topology-based modeling of intrinsically disordered proteins: balancing intrinsic folding and intermolecular interactions. *Proteins*, 79(4), 1251–66.

Gifford, L.K., Carter, L.G., Gabanyi, M.J., Berman, H.M. & Adams, P.D., **2012**. The Protein Structure Initiative Structural Biology Knowledgebase Technology Portal: a structural biology web resource. *Journal of structural and functional genomics*, 13(2), 57–62.

Gillis, A.J., Schuller, A.P. & Skordalakes, E., **2008**. Structure of the Tribolium castaneum telomerase catalytic subunit TERT. *Nature*, 455(7213), 633–7.

Ginalski, K., von Grotthuss, M., Grishin, N. V & Rychlewski, L., **2004**. Detecting distant homology with Meta-BASIC. *Nucleic acids research*, 32, 576–81.

Greider, C.W. & Blackburn, E.H., **1985**. Identification of a specific telomere terminal transferase activity in tetrahymena extracts. *Cell*, 43(2), 405–413.

Grochowski, P. & Trylska, J., **2008**. Continuum molecular electrostatics, salt effects, and counterion binding--a review of the Poisson-Boltzmann theory and its modifications. *Biopolymers*, 89(2), 93–113.

Gront, D., Kmiecik, S. & Kolinski, A., **2007**. Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *Journal of Computational Chemistry*, 28(9), 1593–7.

Gront, D. & Kolinski, A., **2005**. HCPM--program for hierarchical clustering of protein models. *Bioinformatics*, 21(14), 3179–80.

Halperin, I., Ma, B., Wolfson, H. & Nussinov, R., **2002**. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4), 409–43.

Helliwell, J.R., **1984**. Synchrotron X-radiation protein crystallography: instrumentation, methods and applications. *Reports on Progress in Physics*, 47(11), 1403–1497.

Herrmann, T., Güntert, P. & Wüthrich, K., **2002**. Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *Journal of Biomolecular NMR*, 24(3), 171–189.

Horwacik, I., Czaplicki, D., Talarek, K., Kowalczyk, A., Bolesta, E., Kozbor, D. & Rokita, H., **2007**. Selection of novel peptide mimics of the GD2 ganglioside from a constrained phage-displayed peptide library. *International journal of molecular medicine*, 19(5), 829–39.

Horwacik, I., Kurciński, M., Bzowska, M., Kowalczyk, A.K., Czaplicki, D., Koliński, A. & Rokita, H., **2011**. Analysis and optimization of interactions between peptides mimicking the GD2 ganglioside and the monoclonal antibody 14G2a. *International journal of molecular medicine*, 28(1), 47–57.

Huang, Y. & Liu, Z., **2010**. Nonnative interactions in coupled folding and binding processes of intrinsically disordered proteins. *PloS one*, 5(11), 15375.

Itzhaki, L.S., Otzen, D.E. & Fersht, A.R., **1995**. The Structure of the Transition State for Folding of Chymotrypsin Inhibitor 2 Analysed by Protein Engineering Methods: Evidence for a Nucleation-condensation Mechanism for Protein Folding. *Journal of Molecular Biology*, 254(2), 260–288.

Jacobs, S.A., Podell, E.R. & Cech, T.R., **2006**. Crystal structure of the essential N-terminal domain of telomerase reverse transcriptase. *Nature structural & molecular biology*, 13(3), 218–25.

Jamroz, M., Kolinski, A. & Kmiecik, S., **2013**. CABS-flex: server for fast simulation of protein structure fluctuations. *Nucleic acids research*, 11, 1–5.

Janin, J., Henrick, K., Moult, J., Eyck, L. Ten, Sternberg, M.J.E., Vajda, S., Vakser, I. & Wodak, S.J., **2003**. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins*, 52(1), 2–9.

Jiang, F. & Kim, S., **1991**. “Soft docking”: matching of molecular surface cubes. *Journal of Molecular biology*, 219(1), 79–102.

Jones, T.A., Zou, J.Y., Cowan, S.W. & Kjeldgaard, M., **1991**. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallographica Section A Foundations of Crystallography*, 47(2), 110–119.

Kabsch, W. & Sander, C., **1983**. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577–637.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. & Vakser, I.A., **1992**. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences of the United States of America*, 89, 2195–2199.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. & Phillips, D.C., **1958**. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, 181(4610), 662–666.

Kihara, D. & Lu, H., **2001**. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proceedings of the National Academy of Sciences of the United States of America*, 98(18), 10125–10130.

Kinch, L., Yong Shi, S., Cong, Q., Cheng, H., Liao, Y. & Grishin, N. V, **2011**. CASP9 assessment of free modeling target predictions. *Proteins*, 79 Suppl 1, 59–73.

Kolinski, A., **2004**. Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, 51(2), 349–71.

Kolinski, A., Gront, D., Kmiecik, S., Kurcinski, M. & Latek, D., **2006**. Modeling Protein Structure, Dynamics and Thermodynamics with Reduced Representation of Conformational Space. *NIC Workshop 2006, From Computational Biophysics to Systems Biology*, 34, 21–28.

Kolinski, A. & Skolnick, J., **1998**. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. *Proteins*, 32, 475–94.

Kolinski, A. & Skolnick, J., **1994a**. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins*, 18(4), 338–52.

Kolinski, A. & Skolnick, J., **1994b**. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins*, 18(4), 353–66.

Kolinski, A., Skolnick, J. & Yaris, R., **1986**. Monte Carlo simulations on an equilibrium globular protein folding model. *Proceedings of the National Academy of Sciences of the United States of America*, 83, 7267–7271.

Koliński, A. & Bujnicki, J.M., **2005**. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, 61 Suppl 7, 84–90.

Koshland, D.E., **1958**. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2), 98–104.

Kosinski, J., Cymerman, I.A., Feder, M., Kurowski, M.A., Sasin, J.M. & Bujnicki, J.M., **2003**. A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins*, 53 Suppl 6, 369–79.

Krivov, G.G., Shapovalov, M. V & Dunbrack, R.L., **2009**. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, 77(4), 778–95.

Krovat, E., Steindl, T. & Langer, T., **2005**. Recent advances in docking and scoring. *Current Computer-Aided Drug Design*, 1(1), 93–102.

Kumar, A., Wagner, G., Ernst, R.R. & Wuethrich, K., **1981**. Buildup rates of the nuclear Overhauser effect measured by two-dimensional proton magnetic resonance spectroscopy: implications for studies of protein conformation. *Journal of the American Chemical Society*, 103(13), 3654–3658.

Kurcinski, M. & Kolinski, A., **2007a**. Hierarchical modeling of protein interactions. *Journal of Molecular Modeling*, 13(6-7), 691–698.

Kurcinski, M. & Kolinski, A., **2007b**. Steps towards flexible docking: modeling of three-dimensional structures of the nuclear receptors bound with peptide ligands mimicking co-activators' sequences. *The Journal of steroid biochemistry and molecular biology*, 103(3-5), 357–360.

Kurcinski, M. & Kolinski, A., **2010**. Theoretical study of molecular mechanism of binding TRAP220 coactivator to Retinoid X Receptor alpha, activated by 9-cis retinoic acid. *The Journal of steroid biochemistry and molecular biology*, 121(1-2), 124–9.

Leach, A.R., **1994**. Ligand docking to proteins with discrete side-chain flexibility. *Journal of Molecular Biology*, 235(1), 345–356.

Levinthal, C., **1969**. How to fold graciously. In *Mossbauer Spectroscopy in Biological Systems Proceedings of a meeting held at Allerton House Monticello Illinois*. 22–24.

Levitt, M., Sander, C. & Stern, P.S., **1985**. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *Journal of Molecular Biology*, 181(3), 423–447.

Levitt, M. & Warshel, A., **1975**. Computer simulation of protein folding. *Nature*, 253(5494), 694–698.

Lindorff-Larsen, K., Piana, S., Dror, R.O. & Shaw, D.E., **2011**. How fast-folding proteins fold. *Science*, 334(6055), 517–20.

Liwo, A., Kazmierkiewicz, R., Czaplewski, C., Groth, M., Oldziej, S., Wawak, R.J., Rackovsky, S., Pincus, M.R. & Scheraga, H.A., **1998**. United-residue force field for off-lattice protein-structure simulations: III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *Journal of Computational Chemistry*, 19, 259–276.

Małolepsza, E., Boniecki, M., Kolinski, A. & Piela, L., **2005**. Theoretical model of prion propagation: a misfolded protein induces misfolding. *Proceedings of the*

National Academy of Sciences of the United States of America, 102(22), 7835–40.

McPherson, A., **1985**. Diffraction Methods for Biological Macromolecules Part A. *Methods in Enzymology*, 114, 112–120.

Méndez, R., Leplae, R., Lensink, M.F. & Wodak, S.J., **2005**. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 60(2), 150–69.

Metropolis, N., **1987**. The beginning of the Monte Carlo method. *Los Alamos Science*, 15, 125–130.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E., **1953**. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6), 1087.

Metropolis, N. & Ulam, S., **1949**. The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247), 335–341.

Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J. & Weng, Z., **2005**. Protein-Protein Docking Benchmark 2.0: an update. *Proteins*, 60(2), 214–6.

Mitchell, M., Gillis, A., Futahashi, M., Fujiwara, H. & Skordalakes, E., **2010**. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nature structural & molecular biology*, 17(4), 513–8.

Moitessier, N., Englebienne, P., Lee, D., Lawandi, J. & Corbeil, C.R., **2008**. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *British journal of pharmacology*, 153 Suppl, S7–26.

Moore, G.E., **1965**. Cramming more components onto integrated circuits. *Electronics*, 38, 33–35.

Otwinowski, Z. & Minor, W., **1997**. Macromolecular Crystallography Part A. *Methods in Enzymology*, 276, 307–326.

Parker, D., Ferreri, K., Nakajima, T., LaMorte, V.J., Evans, R., Koerber, S.C., Hoeger, C. & Montminy, M.R., **1996**. Phosphorylation of CREB at Ser-133 induces complex formation with CREB-binding protein via a direct mechanism. *Molecular and cellular biology*, 16(2), 694–703.

Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M.S., Davis, F.P., Stuart, A.C., Mirkovic, N., Rossi, A., Marti-Renom, M.A., Fiser, A., Webb, B., Greenblatt, D., Huang, C.C., Ferrin, T.E. & Sali, A., **2004**. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic acids research*, 32, 217–22.

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., van der Spoel, D., Hess, B. & Lindahl, E., **2013**. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7), 845–54.

Qureshi, T. & Goto, N.K., **2012**. Contemporary methods in structure determination of membrane proteins by solution NMR. *Topics in current chemistry*, 326, 123–85.

Radhakrishnan, I., Pérez-Alvarado, G.C., Dyson, H.J. & Wright, P.E., **1998**. Conformational preferences in the Ser133-phosphorylated and non-phosphorylated forms of the kinase inducible transactivation domain of CREB. *FEBS letters*, 430(3), 317–22.

Radhakrishnan, I., Pérez-Alvarado, G.C., Parker, D., Dyson, H.J., Montminy, M.R. & Wright, P.E., **1997**. Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions. *Cell*, 91(6), 741–52.

Rahman, A., **1964**. Correlations in the Motion of Atoms in Liquid Argon. *Physical Review*, 136(2), 405–411.

Rajamani, D., Thiel, S., Vajda, S. & Camacho, C.J., **2004**. Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31), 11287–92.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S. & Baker, D., **2004**. Protein Structure Prediction Using Rosetta. *Methods in Enzymology*, 383, 66–93.

Rouda, S. & Skordalakes, E., **2007**. Structure of the RNA-binding domain of telomerase: implications for RNA recognition and binding. *Structure*, 15(11), 1403–12.

Saibil, J.S. and R.F./ F.R. and H., Klepeis, J.L., Lindorff-Larsen, K., Dror, R.O. & Shaw, D.E., **2009**. Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology*, 19(2), 120–127.

Saibil, J.S. and R.F./ F.R. and H. & Zhang, Y., **2009**. Protein structure prediction: when is it useful? *Current Opinion in Structural Biology*, 19(2), 145–155.

Sánchez, R. & Šali, A., **1997**. Advances in comparative protein-structure modelling. *Current Opinion in Structural Biology*, 7(2), 206–214.

Sander, C. & Schneider, R., **1991**. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1), 56–68.

Shaywitz, a J., Dove, S.L., Kornhauser, J.M., Hochschild, a & Greenberg, M.E., **2000**. Magnitude of the CREB-dependent transcriptional response is determined by the strength of the interaction between the kinase-inducible domain of CREB and the KIX domain of CREB-binding protein. *Molecular and cellular biology*, 20(24), 9409–22.

Sherman, W. & Day, T., **2006**. Novel procedure for modeling ligand/receptor induced fit effects. *Journal of medicinal chemistry*, 49(2), 543–553.

Steczkiewicz, K., Zimmermann, M.T., Kurcinski, M., Lewis, B.A., Dobbs, D., Kloczkowski, A., Jernigan, R.L., Kolinski, A. & Ginalski, K., **2011**. Human telomerase model shows the role of the TEN domain in advancing the double helix for the next polymerization step. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23), 9443–9448.

Stevens, R.C., **2000**. High-throughput protein crystallization. *Current Opinion in Structural Biology*, 10(5), 558–563.

Stevens, R.C., Yokoyama, S. & Wilson, I.A., **2001**. Global efforts in structural genomics. *Science*, 294(5540), 89–92.

Swendsen, R.H. & Wang, J.-S., **1986**. Replica Monte Carlo Simulation of Spin-Glasses. *Physical Review Letters*, 57, 2607–2609.

Tanaka, S. & Scheraga, H.A., **1975**. Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 72(10), 3802–6.

Teichmann, J.D. and J.P./ N.G. and S. & Zhang, Y., **2008**. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3), 342–348.

Tobi, D. & Bahar, I., **2005**. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52), 18908–13.

Trylska, J., Tozzini, V. & McCammon, J.A., **2005**. Exploring Global Motions and Correlations in the Ribosome. *Biophysical Journal*, 89(3), 1455–1463.

Turjanski, A.G., Gutkind, J.S., Best, R.B. & Hummer, G., **2008**. Binding-induced folding of a natively unstructured transcription factor. *PLoS Computational Biology*, 4(4).

Usón, I. & Sheldrick, G.M., **1999**. Advances in direct methods for protein crystallography. *Current Opinion in Structural Biology*, 9(5), 643–648.

Vajda, S., Hall, D.R. & Kozakov, D., **2013**. Sampling and scoring: A marriage made in heaven. *Proteins: Structure, Function, and Bioinformatics*, 81(11), 1874–84.

De Vries, S.J. & Bonvin, A.M.J.J., **2008**. How proteins get in touch: interface prediction in the study of biomolecular complexes. *Current protein & peptide science*, 9(4), 394–406.

Wang, C., Bradley, P. & Baker, D., **2007**. Protein-protein docking with backbone flexibility. *Journal of molecular biology*, 373(2), 503–19.

Wang, J., Cieplak, P. & Kollman, P.A., **2000**. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12), 1049–1074.

Warren, G.L., Andrews, C.W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M.H., Lindvall, M., Nevins, N., Semus, S.F., Senger, S., Tedesco, G., Wall, I.D., Woolven, J.M., Peishoff, C.E. & Head, M.S., **2006**. A critical assessment of docking programs and scoring functions. *Journal of medicinal chemistry*, 49(20), 5912–31.

Warshel, A. & Levitt, M., **1976**. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, 103(2), 227–249.

Wilson, C. & Doniach, S., **1989**. A computer model to dynamically simulate protein folding: studies with crambin. *Proteins*, 6(2), 193–209.

Wodak, S.J., **2007**. From the Mediterranean coast to the shores of Lake Ontario: CAPRI's premiere on the American continent. *Proteins*, 69(4), 697–8.

Wodak, S.J. & Janin, J., **1978**. Computer analysis of protein-protein interaction. *Journal of Molecular Biology*, 124(2), 323–342.

Wuthrich, K., **1969**. High-Resolution Proton Nuclear Magnetic Resonance Spectroscopy of Cytochrome C. *Proceedings of the National Academy of Sciences of the United States of America*, 63(4), 1071–1078.

Yang, L., Song, G. & Jernigan, R.L., **2007**. How Well Can We Understand Large-Scale Protein Motions Using Normal Modes of Elastic Network Models? *Biophysical Journal*, 93(3), 920–929.

Zhang, Y., **2008**. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*, 9(1), 40.

Zhang, Y., Kolinski, A. & Skolnick, J., **2003**. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophysical journal*, 85(2), 1145–1164.

VII Prace stanowiące podstawę rozprawy

Praca I (P.I)

Kurciński, Mateusz and Andrzej Koliński. 2007. **Steps Towards Flexible Docking: Modeling of Three-dimensional Structures of the Nuclear Receptors Bound with Peptide Ligands Mimicking Co-activators' Sequences.** *The Journal of steroid biochemistry and molecular biology* 103(3-5): 357–60.

Praca zawiera wyniki przewidywania trójwymiarowych struktur 10 kompleksów białkowych pomiędzy różnymi białkami z klasy receptorów jądrowych oraz krótkimi peptydami imitującymi wpływ odpowiednich czynników transkrypcyjnych – aktywatorów i represorów. Modelowanie zostało przeprowadzone przy użyciu algorytmu giętkiego dokowania CABS. Otrzymane wyniki wykazały dużą zgodność z danymi krystalograficznymi.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Steroid Biochemistry & Molecular Biology 103 (2007) 357–360
The Journal of Steroid Biochemistry & Molecular Biology
www.elsevier.com/locate/jsbmb

Steps towards flexible docking: Modeling of three-dimensional structures of the nuclear receptors bound with peptide ligands mimicking co-activators' sequences

Mateusz Kurcinski, Andrzej Kolinski *

Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland

Received 30 November 2006

Abstract

We developed a fully flexible docking method that uses a reduced lattice representation of protein molecules, adapted for modeling peptide–protein complexes. The CABS model (Carbon Alpha, Carbon Beta, Side Group) employed here, incorporates three pseudo-atoms per residue—Cα, Cβ and the center of the side group instead of full-atomic protein representation. Force field used by CABS was derived from statistical analysis of non-redundant database of protein structures. Application of our method included modeling of the complexes between various nuclear receptors (NRs) and peptide co-activators, for which three-dimensional structures are known. We tried to rebuild the native state of the complexes, starting from separated components. Accuracy of the best obtained models, calculated as coordinate root-mean-square deviation (cRMSD) between the target and the modeled structures, was under 1 Å, which is competitive with experimental methods, such as crystallography or NMR. Forthcoming modeling study should lead to better understanding of mechanisms of macromolecular assembly and will explain co-activators' effects on receptors activity, especially on vitamin D receptor and other nuclear receptors.

Keywords: Flexible docking; Protein interactions; Nuclear receptor–co-activator complex; Drug design

1. Introduction

Nuclear hormone receptors are ligand-activated transcription factors regulating the expression of target genes and thereby affecting cell reproduction, growth and metabolism [1–3]. Presently, it is believed that human genome contains 48 receptors from this family, but only for half of them ligands have been identified [estrogen receptors (ER), androgen receptor (AR), progesterone receptor (PR), glucocorticoid receptor (GR), mineral corticoid receptor (MR), retinoid X receptors (RXR), retinoic acid receptors (RAR), thyroxine hormone receptors (TR), vitamin D receptor (VDR), peroxisome proliferator-activated receptors (PPAR), liver X receptors (LXR), farnesoid receptor (FXR) and steroid xenobiotic receptor (SXR)]. From the structural point of view all nuclear receptors form a superfamily (according to SCOP [4]) with highly conserved topology of structural motives despite that they bind different ligands. NRs consist of an N-terminal region responsible for ligand-independent transcriptional activation (AF-1), DNA-binding domain (DBD) containing motif of two zinc fingers and C-terminus including ligand-specific binding domain (LBD) and flexible hinge which "locks" ligand upon binding. Usually nuclear receptors are investigated in respect of their interaction only with primary ligands [5–7], but they also form complexes with other molecules. In *holo*-form NRs bind with various cell-specific co-activators, which link receptor with the RNA polymerase II—a gear in transcriptional mechanism, while in *apo*-form NRs form complexes with co-repressors and act as transcriptional suppressors. Abundance of functions of different nuclear receptors and structural similarity between them at the same time makes the NR class very promising pharmacological target.

* Corresponding author. Tel.: +48 22 822 02 11x320; fax: +48 22 822 59 96.
E-mail address: kolinski@chem.uw.edu.pl (A. Kolinski).

0960-0760/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsbmb.2006.12.059

2. Materials and methods

In present work, we describe an application of a recently developed fully flexible docking algorithm to a set of protein complexes from the NR superfamily for which three-dimensional structures are known (so-called 'bound' docking). Our method incorporates the reduced CABS model which has been initially designed for single chain protein folding and performed well in many applications [8]. It has been previously described in great details [9], here we only mention its most basic features.

2.1. Model description

Model CABS is a high resolution reduced model. It assumes protein representation as a three interaction centers per residue (Cα, Cβ, side group). Cα atoms are located on the simple cubic lattice with lattice unit equal to 0.61 Å. Cβ and side group atoms are located off the lattice and their positions are defined by three consecutive Cα atoms. CABS force field was constructed by statistical analysis of non-redundant database of experimentally solved protein structures in the form of histograms reflecting ensembles of various structural properties. It covers most typical interactions such as hydrogen bonds, electrostatics, hydrophobic attraction, but also protein-specific ones: disulfide bridges, centrosymmetrical potential (reflecting the hydrophobic effect), or biases towards regular secondary structure. Sampling of the conformational space is controlled by the Replica Exchange Monte Carlo scheme [10]. Molecules undergo small random conformational changes which are accepted according to the Metropolis Criterion [11]. Additionally, simulation runs in several replicas (copies) in different temperatures and every given amount of simulation steps coordinates of the replicas are exchanged with probability proportional $\exp\{-\Delta E \Delta\beta\}$.

2.2. Protein complexes

We selected from the PDB database 10 high-resolution structures of various nuclear receptor complexes to study the quality of our docking procedure. Starting point for the simulations assumed that both molecules (receptor and co-activator) in their native states were shifted apart from each other to an arbitrary distance of 40 Å between their centers of gravity. Ligand molecules were not explicitly present in simulations, but their influence on the system was incorporated in structural restraints imposed on receptor molecules. None restraints were imposed on co-activators' molecules.

2.3. Clustering and scoring

Trajectories obtained during simulations contained couple of thousands frames, from which only few could be selected as final structures. Since our algorithm uses random search in conformational space, the last frame is not usually the best one. Moreover, the frame with the lowest energy also cannot be simply chosen as the most accurate model, because our method uses statistical force field, where energy does not correlate straightforwardly with the real free energy. In order to select the final structures hierarchical clustering was applied to every trajectory. We used the HCPM [12] program, with the single-link clustering procedure and coordinate root-mean-square deviation (cRMSD) as a similarity measure. Structure of the best representative of the cluster was selected as a frame closest (in the means of cRMSD) to the centroid of the cluster.

Except for the minimal cRMSD, calculated for the whole complex, the quality of the predicted complex structures was assessed in terms of the following measures [13]:

Ligand cRMSD, defined as the cRMSD calculated only for the ligand backbone after superimposition of the receptor structures.

Fraction of native contacts, defined as the number of correct contacts between receptor's and ligand's residues in the predicted complex, divided by the number of contacts in the native structure of the complex. A pair of residues is arbitrary considered to be in contact if distance between their Cα atoms is less then 10 Å.

3. Results and discussion

Final models were selected from trajectories in two different manners. At first every frame of the trajectory was compared with the native structure from the PDB database in

Table 1
Best frames selected from the trajectories compared with crystallographic structure

PDB code	Chains	NR class	Minimal cRMSD (Å)	Ligand cRMSD (Å)	Fraction of native contacts
1KKQ	A E	PPARα	4.26	14.88	0.11
1KV6	A C	ERR	1.48	1.71	0.68
1M2Z	A B	GR	2.99	8.46	0.18
1MVC	A B	RXRα	0.49	1.20	0.94
1NQ7	A B	RORβ	0.42	0.86	0.90
1NRL	A C	PXR	5.09	22.64	0.00
1OSV	B D	FXR	1.37	3.94	0.36
1RJK	A C	VDR	0.58	1.47	0.63
1XB7	A P	ERRα	7.33	19.09	0.02
3ERD	A C	ERα	0.87	2.03	0.64

Table 2
Final models selected in clustering procedure, compared with crystallographic structures

PDB code	Cluster	Minimal cRMSD (Å)	Ligand cRMSD (Å)	Fraction of native contacts
1KKQ	1	5.76	15.20	0.09
	2	8.71	29.67	0.00
	3	**4.99**	**17.14**	**0.15**
1KV6	**1**	**2.29**	**4.24**	**0.58**
	2	8.34	40.33	0.00
	3	7.40	35.50	0.00
1M2Z	1	10.60	39.30	0.00
	2	9.82	36.37	0.00
	3	**9.08**	**33.87**	**0.00**
1MVC	1	0.95	3.86	0.70
	2	0.83	3.19	0.62
	3	**0.74**	**2.78**	**0.74**
1NQ7	**1**	**0.47**	**0.89**	**0.90**
	2	0.70	2.84	0.74
	3	3.71	19.45	0.06
1NRL	**1**	**5.97**	**26.41**	**0.00**
	2	6.24	27.59	0.00
	3	6.43	28.24	0.00
1OSV	**1**	**1.70**	**4.96**	**0.31**
	2	4.91	21.93	0.00
	3	5.76	26.37	0.00
1RJK	**1**	**0.78**	**3.08**	**0.72**
	2	2.33	10.93	0.30
	3	2.07	9.60	0.32
1XB7	1	10.46	40.28	0.00
	2	8.08	25.50	0.00
	3	**7.96**	**24.37**	**0.00**
3ERD	1	6.61	33.07	0.00
	2	**1.72**	**7.41**	**0.23**
	3	6.99	34.98	0.00

Best models are presented in bold.

the means described in the previous paragraph and the final model was chosen as the frame with the lowest cRMSD to the native structure. Results are presented in Table 1. Afterwards we applied different method of selecting final models from the trajectory, referring to the cases when native structure of the complex is unknown and selection of the final model must be made using non-comparative techniques. Clustering procedure was applied to trajectory and final models were chosen as representative structures from the three biggest clusters (containing the largest number of frames). Quality of obtained models was assessed by comparing them with the native structure. For five models the top-scored cluster was at the same time the biggest one. In remaining cases it was among the three biggest. Results are presented in Table 2.

In three cases (1MVC, 1NQ7, 1RJK) obtained models may be considered as very good (cRMSDs < 1 Å), in other three (1KV6, 1OSV, 3ERD) as good (cRMSD < 2.5 Å) (Figs. 1 and 2). The remaining four models are inaccurate (cRMSD ≈ 5–10 Å), however two of them (1KKQ, 1NRL) still may be a source of qualitative structural information about the location of the binding site, since in these cases co-activators were docked in correct site on the surface of the receptor, but were wrongly oriented. It is worth mentioning that no additional information about location of the binding site was used in modeling. It indicates that our algorithm is able to utilize structural information contained in receptor's

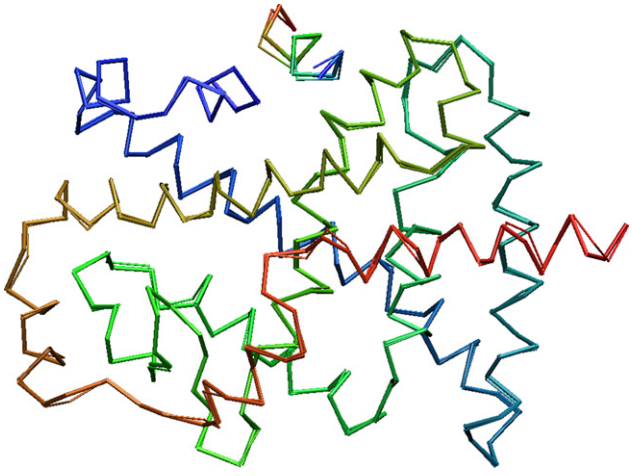


Fig. 1. Best model (thick lines) of 1NQ7 in the alpha carbon representation superimposed onto crystallographic structure (thin lines).

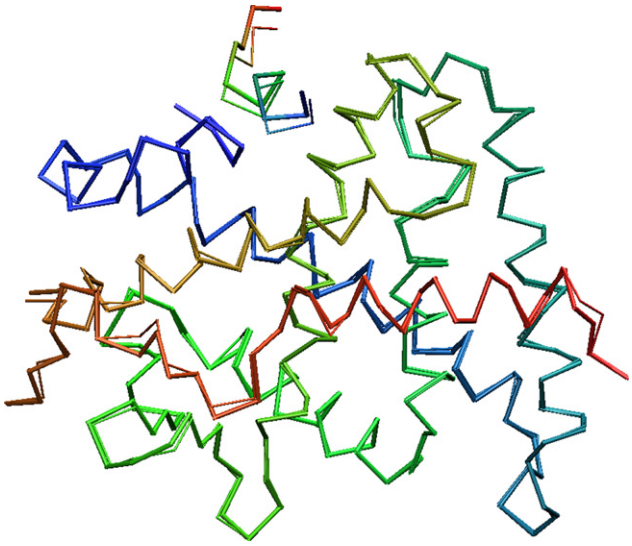


Fig. 2. Best model (thick lines) of 1RJK in the alpha carbon representation superimposed onto crystallographic structure (thin lines).

structure without calculating any sort of molecular surface, but only by efficient exploration of the energy landscape.

We discovered that for those complexes that were modeled with the highest accuracy the biggest cluster contained almost half of the trajectory frames, while for those worse modeled only about 20% frames belonged to the first cluster. This may be a strong premise to verify if the model is correct in 'unbound' docking cases, where no native structure to compare with is available. Here we present application of our algorithm only to 10 examples of protein complexes, which are additionally close homologues. In order to verify above mentioned thesis, wider experiment must be conducted, which would cover at least couple of tens various proteins.

4. Conclusion

We developed a new algorithm for fully flexible docking of peptides and proteins. It is based on previously described CABS mesoscopic modeling tool which was successfully used for study of protein dynamics and thermodynamics [14,15] and prediction of protein structures [16]. It has been shown here that the new modeling tool is already capable of producing correct high-resolution structures of protein–peptide complexes. This should be very important for understanding of protein interactions, signaling pathways and computed aided design of new drugs. Work in progress aims on designing of a complete protein interactions modeling tool, which could be used to simulate small ligand docking, macromolecular assembly of protein domains and interactions between proteins and nucleic acids.

Acknowledgment

Computational part of this work was done using the computer cluster at the Computing Center of Faculty of Chemistry, Warsaw University.

References

[1] R.M. Evans, The steroid and thyroid hormone receptor superfamily, Science 240 (4854) (1988) 889–895.
[2] D.J. Mangelsdorf, et al., The nuclear receptor superfamily: the second decade, Cell 83 (6) (1995) 835–839.
[3] A. Chawla, et al., Nuclear receptors and lipid physiology: opening the X-files, Science 294 (5548) (2001) 1866–1870.
[4] A.G. Murzin, et al., SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (4) (1995) 536–540.
[5] P. Rotkiewicz, et al., Model of three-dimensional structure of vitamin D receptor and its binding mechanism with 1alpha,25-dihydroxyvitamin D(3), Proteins 44 (3) (2001) 188–199.
[6] R.R. Sicinski, et al., 2-Ethyl and 2-ethylidene analogues of 1alpha,25-dihydroxy-19-norvitamin D(3): synthesis, conformational analysis, biological activities, and docking to the modeled rVDR ligand binding domain, J. Med. Chem. 45 (16) (2002) 3366–3380.
[7] W. Sicinska, P. Rotkiewicz, H.F. DeLuca, Model of three-dimensional structure of VDR bound with Vitamin D3 analogs substituted at carbon-2, J. Steroid Biochem. Mol. Biol. 89–90 (1–5) (2004) 107–110.
[8] A. Kolinski, J.M. Bujnicki, Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models, Proteins 61 (Suppl. 7) (2005) 84–90.
[9] A. Kolinski, Protein modeling and structure prediction with a reduced representation, Acta Biochim. Pol. 51 (2) (2004) 349–371.
[10] R.H. Swendsen, J.S. Wang, Replica Monte Carlo simulation of spin glasses, Phys. Rev. Lett. 57 (21) (1986) 2607–2609.
[11] N. Metropolis, et al., Equation of state calculations by fast computing machines, J. Chem. Phys. 51 (1953) 1087–1092.
[12] D. Gront, A. Kolinski, HCPM–program for hierarchical clustering of protein models, Bioinformatics 21 (14) (2005) 3179–3180.
[13] S. Vajda, C.J. Camacho, Protein-protein docking: is the glass half-full or half-empty? Trends Biotechnol. 22 (3) (2004) 110–116.
[14] D. Ekonomiuk, M. Kielbasinski, A. Kolinski, Protein modeling with reduced representation: statistical potentials and protein folding mechanism, Acta Biochim. Pol. 52 (4) (2005) 741–748.
[15] S. Kmiecik, et al., Denatured proteins and early folding intermediates simulated in a reduced conformational space, Acta Biochim. Pol. 53 (1) (2006) 131–144.
[16] M. Boniecki, et al., Protein fragment reconstruction using various modeling techniques, J. Comput. Aided Mol. Des. 17 (11) (2003) 725–738.

Praca II (P.II)

Kurciński, Mateusz and Andrzej Koliński. 2007. **Hierarchical Modeling of Protein Interactions.** *Journal of Molecular Modeling* 13(6-7): 691–98.

W pracy przedstawiono schemat wielostopniowej procedury w pełni giętkiego dokowania. Zaprezentowano wyniki dla 10 testowych przypadków, obejmujących zarówno układy typu białko – peptyd, jak również kompleksy homodimerów. Otrzymano w większości poprawne modele o akceptowalnej rozdzielczości.

J Mol Model (2007) 13:691–698
DOI 10.1007/s00894-007-0177-8

ORIGINAL PAPER

Hierarchical modeling of protein interactions

Mateusz Kurcinski · Andrzej Kolinski

Received: 15 November 2006 / Accepted: 18 January 2007 / Published online: 13 February 2007

Abstract A novel approach to hierarchical peptide–protein and protein–protein docking is described and evaluated. Modeling procedure starts from a reduced space representation of proteins and peptides. Polypeptide chains are represented by strings of alpha-carbon beads restricted to a fine-mesh cubic lattice. Side chains are represented by up to two centers of interactions, corresponding to beta-carbons and the centers of mass of the remaining portions of the side groups, respectively. Additional pseudoatoms are located in the centers of the virtual bonds connecting consecutive alpha carbons. These pseudoatoms support a model of main-chain hydrogen bonds. Docking starts from a collection of random configurations of modeled molecules. Interacting molecules are flexible; however, higher accuracy models are obtained when the conformational freedom of one (the larger one) of the assembling molecules is limited by a set of weak distance restraints extracted from the experimental (or theoretically predicted) structures. Sampling is done by means of Replica Exchange Monte Carlo method. Afterwards, the set of obtained structures is subject to a hierarchical clustering. Then, the centroids of the resulting clusters are used as scaffolds for the reconstruction of the atomic details. Finally, the all-atom models are energy minimized and scored using classical tools of molecular mechanics. The method is tested on a set of macromolecular assemblies consisting of proteins and peptides. It is demonstrated that the proposed approach to the flexible docking could be successfully applied to prediction of protein–peptide and protein–protein interactions. The obtained models are almost always qualitatively correct, although usually of relatively low (or moderate) resolution. In spite of this limitation, the proposed method opens new possibilities of computational studies of macromolecular recognition and mechanisms of assembly of macromolecular complexes.

Keywords Protein modeling · Flexible docking · Macromolecular recognition · Protein assemblies · Monte Carlo

M. Kurcinski · A. Kolinski (✉)
Faculty of Chemistry,
Warsaw University,
ul. Pasteura 1,
02-093 Warsaw, Poland
e-mail: kolinski@chem.uw.edu.pl

Introduction

Tens of thousands of experimental structures of proteins are known already. Although this is only a small fraction (range of 0.1%) of the number of known protein sequences, it is safe to assume that the known structures are representative for at least half of all existing protein structures. Indeed, using advanced contemporary methods of comparative modeling, it is possible to build reasonable quality theoretical molecular models for more than 50% of newly sequenced proteins [1]. This opens enormous opportunities for theoretical interpretation of protein biological functions at the molecular level. Modeling of enzymatic activity, mechanisms of molecular signaling, protein associations (including autocatalytic changes of native structures, as the ones responsible for some neurodegenerative diseases), and other processes has recently become possible. At the same time, computer studies of biomacromolecular processes face numerous technical difficulties, due to the high level of molecular complexity of the systems under consideration.

One of the most important and challenging tasks of computational molecular biology is prediction of interactions between biologically important molecules. Docking of

smaller molecules into receptors and protein–protein docking are essential for computer-aided, rational drug design, understanding metabolic pathways and theoretical interpretations of molecular basis of neurodegenerative diseases, etc. In general, the term "docking" means finding the lowest free energy conformation(s) of a ligand-receptor complex [2]. The traditional approach assumes knowledge of the structure of the macromolecular components [3, 4]. Thus, in the case of small molecule docking, its location in respect to the receptor and its internal conformation, compatible with the complex structure and intra-and intermolecular interactions, need to be calculated. In the cases of protein–protein docking, the two structures need to be connected, i.e., the interface(s) of the complex has to be found. So called "flexible docking" is usually limited to an optimization of the ligand structure, relaxation of the involved side chains, and (sometimes) very limited readjustment of portions of the main chain of proteins [5, 6]. This kind of approach sometimes works very well. However, in general, the assumptions of the traditional docking approaches could be non-physical. Structures of components of an assembly are always a function of the interactions in the entire complex. Frequently, the same molecule (macromolecule) in isolation has a qualitatively different structure from that observed in various complexes. There are numerous examples where, upon a ligand docking, the structure of the receptor changes dramatically. The same is true the for protein–protein association.

In this work, we propose a new approach, which is the first step towards a fully flexible docking. At this moment, the new method is limited to docking of peptides to proteins and protein–protein docking only, with some restrictions imposed on the range of conformational changes of the macromolecular components, induced by the assembly process. The approach is hierarchical—a multiscale molecular modeling is employed. Here, examples of two-molecule complexes are discussed and, although the algorithms are general, larger than two-molecule systems could be treated in the same way. The modeling process starts from a random mutual orientation of the molecules under consideration and a random conformation of one of them (the smaller one or the one randomly selected if the two components are of the same molecular mass). Simulation of the assembly process is done using the multichain version of our reduced-space protein modeling tool CABS [7]. CABS (CA, Cb, and Side chain) employs a united atom representation of protein conformations. A residue in a model polypeptide chain is represented by up to four interaction centers corresponding to the alpha carbon, the beta carbon, the center of mass of the remaining portion of the side chain (where applicable) and pseudoatoms located in the center of the virtual bonds connecting the alpha carbons. Conformational space of this model is sampled by means of multicopy Monte Carlo method (REMC, Replica Exchange Monte Carlo) versions of which are also known as the Simulated Tempering MC [8]. Subsequently, multiple structures resulting from the REMC simulations with the CABS model are subject to clustering, rebuilding the atomic details, all-atom optimization/minimization, and scoring of the models. Finally, the model structures are compared with the known crystallographic structures. Most of the modeled assemblies were very close to the native ones, showing the predicting power of the new method.

Methods

Figure 1 shows a flow-chart of the flexible docking procedure employed in this work. The simulations have a

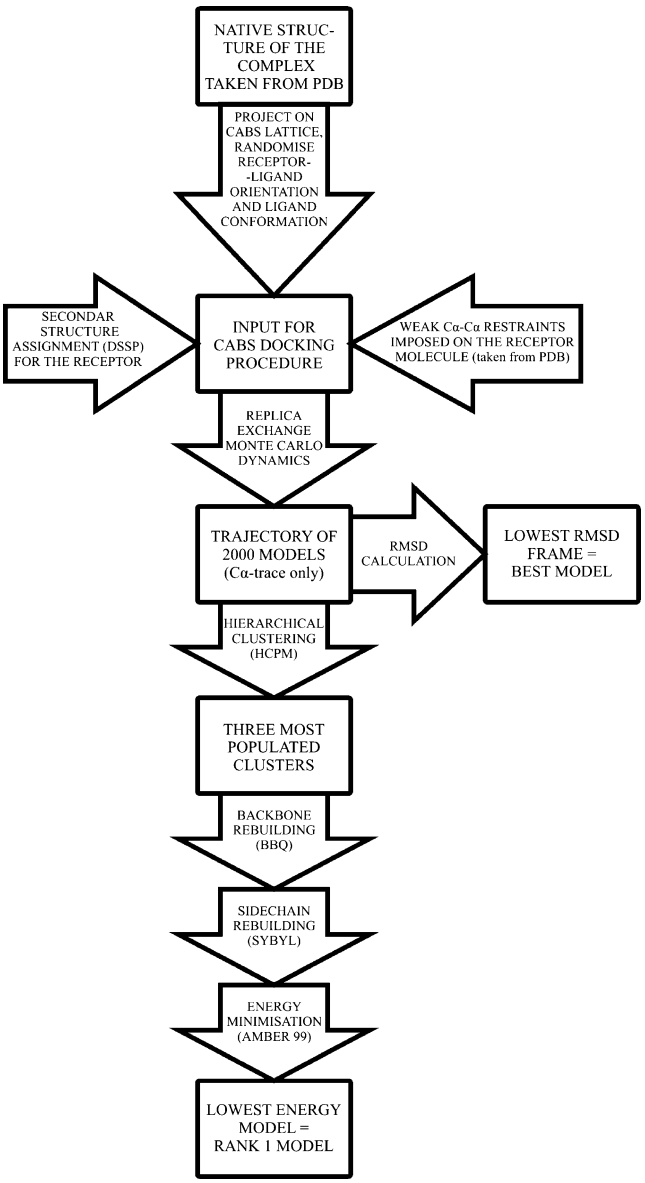


Fig. 1 Flow chart of multiscale assembly of protein complexes

multiscale character. Search of the conformational space is done by means of REMC simulations with CABS reduced model of polypeptides. Refinement and final selection of molecular models is done on the all-atom level using an AMBER 99 [9] force field. Details of the proposed methodology are given below.

CABS model and simulation method

CABS is a simplified model of polypeptides. A detailed description of this model and its force field has recently been published [7]. The model has been extensively tested during the CASP6 (Critical Assessment of Protein Structure Prediction) community-wide experiment [10]. The results of CASP6 have proved that CABS was one of the two best protein folding algorithms, especially when applied to more difficult cases of blind structure prediction [11, 12]. Other applications of CABS, as structure prediction based on sparse NMR data, modeling of folding dynamics and thermodynamics, etc., have also been described recently [13–15]. Here, for reader convenience, we provide just a concise outline of the CABS features.

A single residue in the CABS model is represented by up to four interaction centers corresponding to alpha carbons, beta carbons, the center of mass of the remaining portion of the side chain (where applicable) and a pseudoatom located in the center of the virtual Ca-Ca bond, corresponding roughly to the center of mass of the peptide bond. The positions of alpha carbons are restricted to the underlying cubic lattice, with the mesh size equal to 0.61Å. The lengths of the Ca-Ca virtual bonds are allowed to oscillate around their physical value of 3.78Å. These pseudobonds belong to a set of 800 lattice vectors. Thus, possible lattice artifacts can be safely avoided. The fluctuating bond length facilitates very fast sampling and a better packing of the model side groups. Lattice-restricted Ca-trace provides a convenient reference frame for the definition of the (off-lattice) coordinates of the remaining united atoms. Two-rotamers approximation of the side groups' mobility is assumed. Geometry of the model amino acids is derived from statistical analysis of high resolution protein structures. Lattice-based reference frame of the model chain facilitates very rapid and straightforward calculations of conformational transitions.

The interaction scheme of CABS consists of several heuristic potentials derived from the statistical analysis of the structural regularities seen in the folded proteins. The short range potentials have the form of energy histograms as functions of the distances between i-th and i+2, i+3, and i+4th alpha carbons, the amino acid composition and the chirality of the corresponding fragments. The long range potentials consist of hard core repulsions between the Ca, Cb united atoms and peptide bond pseudoatoms, and soft (finite) repulsions between the side chains. Attractive parts of the side chain pairwise potentials have a form of square-well potentials. The strength of the side chain interactions is context-dependent and is different for different orientations of the interacting groups and for different types of the local geometry of the main chain. Additionally, there is a model of the main chain hydrogen bonds defined as a set of geometrical criteria for a configuration of the interacting fragments of the main chain that is consistent with the presence of the main chain hydrogen bonds.

Intermolecular interactions and intramolecular interactions were treated in the same way. This is certainly an approximation. There are, however, two reasons partially justifying this approximation. First, a derivation of separate statistical potentials for the interfaces is more difficult due to the limited set of the crystallographic data available. Second, as it is shown in this work, the assumption of the same potentials works very well, indicating that the possible differences in these interactions are probably not that significant. Nevertheless, this issue needs to be further investigated.

Conformational updating of the CABS chains consists of a random sequence of one residue, two, three and four residues local micromodifications. For instance, the one residue update requires a small random displacement of a randomly selected alpha carbon and subsequent rearrangements of the three involved side chains. Occasionally, a small displacement of a larger fragment of a chain (up to 22 residues) is attempted, allowing for a rigid-body component of the simulated motion. The sequence (types of motion and their placement along the polypeptide chain) is controlled by a random algorithm. As a result, the entire chain can relax its internal coordinates and move in space. Consequently, the ligand is allowed to sample the full spectrum of its internal coordinates and all possible locations in respect to the receptor. The receptor mobility is limited to a local relaxation (backbone and side chains) by a set of weak harmonic restraints imposed onto a small fraction of the Ca-Ca distances in the receptor (five restraints per residue). The strength of these restraints was sufficiently low to allow 2–4 Å relaxation of the receptor starting structure. The last limitation (receptor's restraints) can be omitted, although it would lead to significantly poorer quality of the resulting models. In future work, it is planed to identify rigid and flexible regions of the receptor, thereby allowing docking with large scale structural rearrangements of the all assembling components. The details of Monte Carlo dynamics employed here are described in detail elsewhere [7]. Thanks to the lattice-type representation of the model chains, the computation of the attempted micromodifications is extremely simple and requires only a few references to randomly selected elements of large tables describing precomputed, allowed conformational transitions. Thus, the

simulations with CABS are much faster than would be the case with an otherwise similarly reduced model in the continuous space.

Sampling of the CABS model is done by means of a version of the Replica Exchange Monte Carlo method. The starting conformations were constructed as follows. First, the receptor (or one of the members of a homodimer structure) has been projected onto the CABS representation using the native structure from PDB as a template. Then, the second molecule was added in a random conformation and in a random position in respect to the first molecule. Each replica had a different random configuration. During the simulations, both molecules were mobile, although the structure of the first molecule (receptor) was controlled by a set of weak restraints, taken from the native structure of the assembly. The strength of these restraints was high enough to keep the receptor structure in a region of globally native-like conformations, though allowing for significant readjustments of the assembly interface. Additionally, a weak harmonic force has been turn-on, when the two molecules separated on distances precluding any chain contacts. This was done for a purely technical reason, namely to avoid useless sampling of both molecules in separation. At distances allowing even sparse contacts of two simulated molecules the harmonic force has been turn-off.

The results of single simulations were stored as pseudotrajectories containing 2,000 snapshots of the system Ca-traces. Depending on the size of studied assemblies simulations took a couple of hours (for the smallest assembly) to a couple of days (for the largest one studied here) on a single LINUX box.

Clustering

Simulation trajectories were subject to a hierarchical clustering procedure using the HCPM software [16] (http://www.biocomp.chem.uw.edu.pl/HCPM). The clustering was stopped when about 50% of the structures were assigned to the growing clusters. The structures closest to the clusters' centroids of the three biggest clusters were selected for further processing. For the purpose of an additional evaluation of the subsequent model-ranking procedure, one more structure (the fourth one), which was the closest to the native (RMSD-coordinate root-mean-square deviation for the alpha traces after the best superimposition with the crystallographic structure) in the entire trajectory of 2,000 snapshots, was also selected.

All-atom refinement and model selection

Selected reduced structures from the Monte Carlo simulations (centroids of the top three clusters and the best structure observed in a simulation) were used as scaffolds for the all-atom building. The recently developed BBQ (Backbone Building from Quadrilaterals) program was first applied to the reconstruction of the all-backbone atoms [17] (http://www.biocomp.chem.uw.edu.pl/BBQ). The BBQ algorithm employs a large set of high resolution protein fragments which are superimposed on the Ca-trace. The algorithm is very fast, accurate and tolerates the inaccurate distances between alpha carbons from CABS simulations very well. In the next step, approximate positions of the side chains are calculated using a large library of side chain rotamers and the main chain coordinates as the reference frame [18].

The crude all-atom models were finally refined during a relatively short minimization with an AMBER force field using TRIPOS software [9]. The solvent was treated in the implicit fashion using the Generalized Born solvatation model. It has been proved that the energy from the all-atom minimization correlates better with the RMSD than does the CABS energy. Thus, the multiscale approach improves the accuracy of the model selection. A similar effect was also observed by Baker [19] for a combination of the Rosetta reduced models with a refinement on the atomic level.

Test set of macromolecular assemblies

A set of 11 protein–peptide and protein–protein assemblies was selected for the test predictions using the method described above. Although relatively small, the test set seemed be representative. It contained heterodimers and homodimers, the size of the component of the assemblies ranging from 5 to 197 residues, and the participating proteins representing various structural classes (alpha, beta and alpha/beta). An additional criterion of the selection was high resolution of the crystallographic structures, which could be of some importance for a proper evaluation of the accuracy of predictions. Table 1 contains a brief description of the studied systems.

Results and discussion

Using the pipeline described in Methods and outlined in the flow-chart given in Fig. 1, molecular models for the 11 test systems (Table 1) were generated and evaluated. The reduced conformational space modeling with the CABS is the core element of this pipeline. The CABS modeling started from a number of random configurations of the two molecules–replicas in the REMC simulations. The conformations of the first molecule (receptor, or an arbitrarily selected molecule from a homodimer) were weakly restrained to the near-native region of its conformational space. The second molecule was fully mobile in a neighborhood of the first one and internally flexible during

Table 1 Summary of the test set of macromolecular assemblies

PDB code	Complex	Receptor class	Receptor length	Ligand length
1A2X	Troponin C – Troponin I	α	158	31
1CKA	CRK SH3 Domain – Proline rich peptide	β	57	9
1KLQ	MAD2 Protein – MAD2 binding peptide	α/β	197	10
1KYF	Alpha adaptin C – Growth factor substrate	β	117	5
1NTV	Disabled PTB domain – ApoER2 peptide	β	152	10
1OGT	HLA antigen – Vasoactive polypeptide	α/β	179	9
1OKH	Crambin-like homodimer	α/β	46	46
1RPR	ROP homodimer	α	63	63
1VWA	Streptavidin – Peptide containing HPQ	β	121	7
2BBM	Calmodulin – Myosin light chain	α	148	26
2ZTA	GCN4 leucine zipper	α	31	31

the simulations. Totals of 2,000 snapshots were extracted from each simulation, at equal intervals of the clock of the simulation algorithm.

After clustering, the centroids of the top three clusters were subject to the all-atom reconstruction and refinement, and the lowest energy (AMBER 99) model was reported as the RANK 1 model in Table 2. In order to also evaluate the proposed strategy for the selection of models the best CABS structure observed in the simulations was all-atom reconstructed and energy minimized in exactly the same

Table 2 Summary of the modeling results

PDB code	Model	AMBER99 energy	Cluster size	RMSD			Fraction of native contacts
				Total	Receptor	Ligand	
1A2X	RANK 1	−3,086.64	957	1.76	1.02	3.83	0.66
	BEST	−2,562.66	1	1.24	0.95	2.24	0.70
1CKA	RANK 1	−852.46	489	2.69	0.77	7.87	0.57
	BEST	−852.71	186	1.60	0.50	4.77	0.64
1KLQ	RANK 1	−968.02	157	1.56	1.04	5.58	0.75
	BEST	−1,001.23	712	1.30	0.97	4.10	0.74
1KYF	RANK 1	−1,931.22	21	1.87	1.07	7.80	0.57
	BEST	−1,573.20	1	1.37	1.13	4.05	0.52
1NTV	RANK 1	−2,456.88	117	2.20	0.99	8.17	0.58
	BEST	−2,337.50	362	1.09	0.87	2.77	0.65
1OGT	RANK 1	−3,261.38	19	1.92	0.55	8.65	0.30
	BEST	−3,612.17	1	1.18	0.47	5.08	0.51
1OKH	RANK 1	−958.41	720	3.63	1.12	9.27	0.48
	BEST	−437.67	720	2.14	1.29	4.01	0.59
1RPR	RANK 1	−2,050.36	542	5.21	1.68	10.93	0.40
	BEST	−2,321.46	542	4.95	1.21	10.57	0.45
1VWA	RANK 1	−1,729.21	28	2.95	1.32	11.54	0.30
	BEST	−2,339.11	897	1.59	1.20	4.70	0.38
2BBM	RANK 1	−3,030.26	640	2.30	0.91	5.61	0.80
	BEST	−2,872.88	1	1.06	0.55	2.49	0.78
2ZTA	RANK 1	−1,400.24	687	1.69	1.18	2.69	0.51
	BEST	−1,148.31	3	1.43	1.35	2.66	0.58

RANK 1 the lowest energy (AMBER 99) all-atom model of a cluster centroid from the CABS simulations
BEST: the all-atom model of the best structure from the set of 2000 CABS models
Total coordinate RMSD of the entire complex after the best superimposition with the crystallographic structure
Receptor coordinate RMSD for the receptor (or the first chain in a homodimer) after the best superimposition with the crystallographic structure of the receptor alone
Ligand coordinate RMSD for the ligand (or the second chain in the homodimer) after the best superimposition of the receptor (or the first chain of the homodimer)
Fraction of native contacts fractions of native contacts observed in the model structures on their protein-protein (or protein-peptide) interfaces. After Vajda [20] the "contacts" are defined for pairs of alpha-carbon atoms using 10Å cut-off

fashion as was done for the clusters' centroids. These are reported as the BEST models in Table 2.

Various measures of the accuracy of the models could be used. In Table 2, we reported the coordinate root-mean-square deviation (RMSD) for the entire complex after the best superimposition (the column with the header Total), the RMSD for the restrained part of the system (Receptor), the RMSD for the entirely free-moving part of the system after the best superimposition of the restrained part of the assembly (Ligand), and the fraction of the native contacts observed in the complex interface.

Analysis of the modeling results summarized in Table 2 leads to a number of interesting observations. First of all, it can be noted that in all cases the resulting models are qualitatively correct (see also Figs. 2, 3, 4 and 5), as evidenced by RMSD values for entire complexes (ranging from 1.06 Å for the 2BBM, Calmodulin-Myosin light chain complex, to 5.21 Å in the worst case of 1RPR, ROP homodimer, with a typical value below 2 Å for most cases) and significant fractions of recovered native contacts in the model interfaces (typically above 50%). It should be pointed out that the employed definition of contacts [20] is very sensitive to small structural changes and therefore the value of 50% is highly significant and reflects qualitatively correct association interface. Even in the worst case of the ROP homodimer, the high value of RMSD (5.21 Å) does not mean a qualitatively wrong prediction. Figure 5 shows that the two molecules are in a proper mutual orientation and that the high value of RMSD is caused by misfolding of small terminal fragments of one of the monomers.

Very small values of the first molecules' RMSD (typically below 1 Å) indicate that the CABS model, even with relatively weak distance restraints, facilitates a very accurate representation of protein conformations, in spite of its reduced representation. Indeed, during the CASP6 experiment several comparative models built by means of the template restrained CABS simulations were of an experimental quality (in the range of 1.5 Å RMSD from the native). In some cases, the models were better than the used templates [10] (http://www.predictioncenter.org/casp6). On the other hand, such good quality of the restrained parts of the complexes studied in this work does

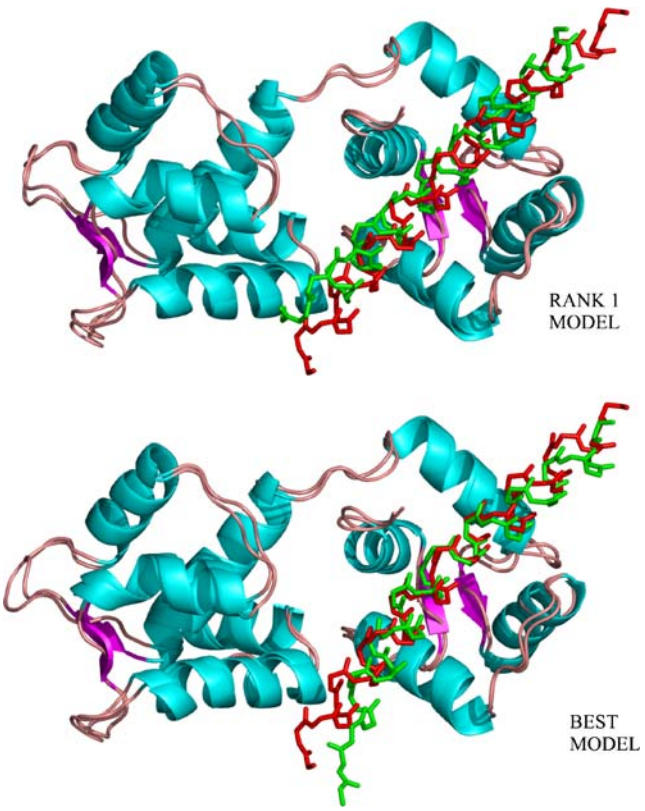


Fig. 2 Superimposition of the predicted and native structures of 1A2X complex. The receptor is shown as ribbons. Covalent structure for heavy atoms is shown for the peptide ligand. *Rank 1 model* is the lowest energy (according to the AMBER force-field) model after refinement of the clusters' centroids. The *best model* is the refined all-atom model built on the best (according to the RMSD deviation from the crystallographic structure) CABS structure found in the trajectory of 2000 snapshots. Ligand native structure is shown in red, modeled in green

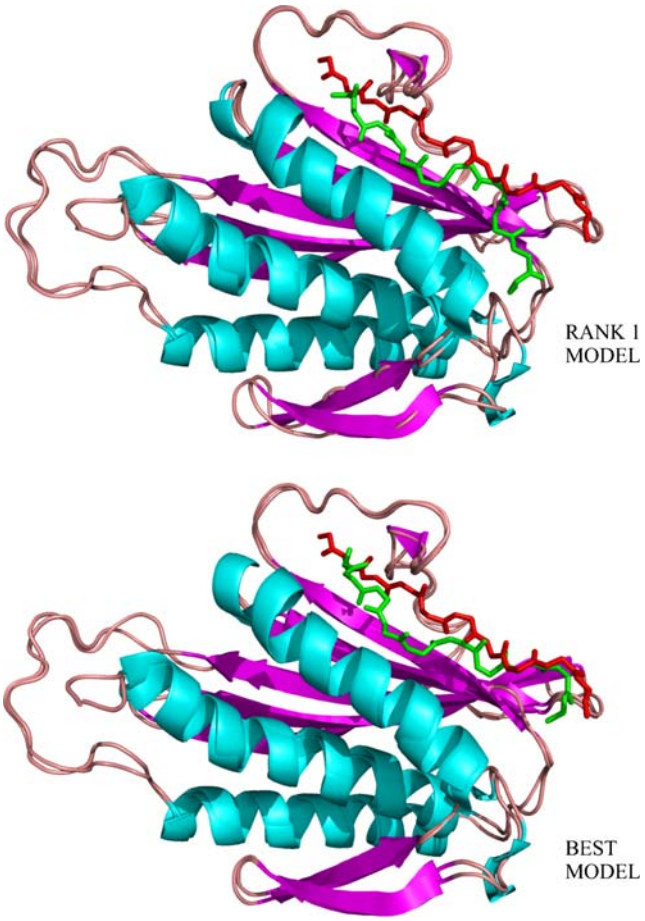


Fig. 3 Superimposition of the predicted and native structures of 1KLQ complex. The receptor is shown as ribbons. Covalent structure for heavy atoms is shown for the peptide ligand. See also the legend to Fig. 2

Fig. 4 Superimposition of the predicted and native structures of 1OKH homodimer. Native structure in cyan and predicted in magenta ribbons, respectively

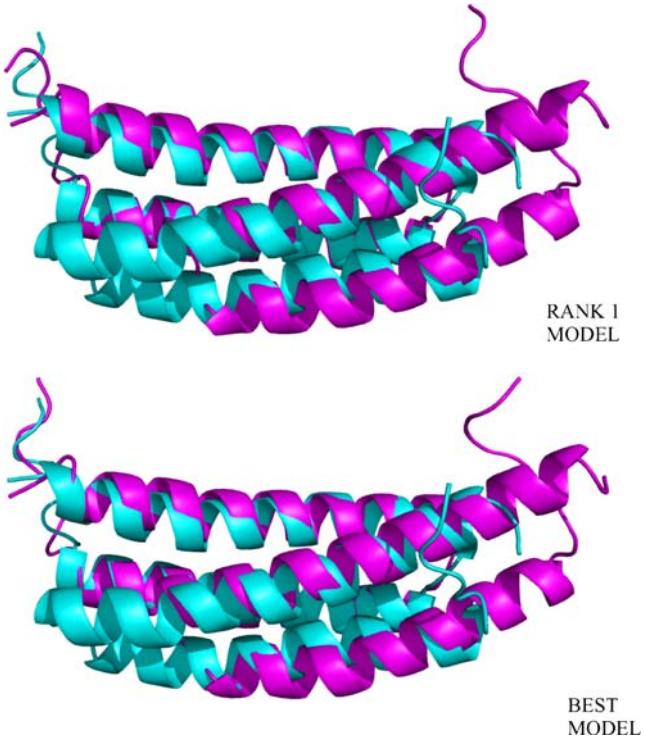


Fig. 5 Superimposition of the predicted and native structures of 1RPR homodimer. Native structure in cyan and predicted in magenta ribbons, respectively

not prove that a completely de novo macromolecular assembly using the CABS model is feasible. This problem is beyond the scope of this work. Previous work has shown that for not too large systems an unrestrained assembly is possible, although the accuracy of the resulting models would be lower. Incomplete templates for multimers (as also demonstrated during CASP6) are often sufficient for a successful modeling by CABS.

The stringiest criterion of the model quality is the RMSD of the ligand (or the second molecule of a dimer) after the best superimposition of the receptor structure (Ligand). Here, the deviations from the crystallographic structure are of the largest magnitude (ranging from 2 to 11 Å). Nevertheless, in all cases the native docking interface was at least partially recovered during modeling. Small structural inaccuracies of the internal conformations of the ligands are the main sources of the RMSD errors (see Figs. 2, 3, 4 and 5). Good examples of such a situation provide the models of ROP homodimers (see Fig. 5).

Finally, the proposed procedure for the selection of the best models needs a comment. The all-atom refinement certainly leads on average to a better selection than that based on ranking according to the CABS energy or/and based on the clusters' size after the CABS simulations (data for the CABS-only modeling not given). The proposed model selection procedure almost always (the case of 1VWA is an exception) leads to reasonable quality models. At the same time (as demonstrated in Table 2) better models could always be found in the simulation trajectories. Thus, there is quite a lot of room for an improvement of the proposed modeling method just by a design of more dependable model scoring and selection procedures. It is evident (Table 2) that the AMBER energy after minimization is not always the lowest for the best structures. Other force fields and more elaborated all-atom refinement protocols need to be investigated. Model selection is always an important component of all approaches to reduced-space modeling of proteins. It has been given significant attention during the CASP experiments [10], and it has been concluded that a fully satisfying solution to the problem remains to be found. The most promising possibilities for an improvement of this aspect of proteins and macromolecular assembly modeling include: analysis of a larger number of clusters, modification of the clustering criteria (not only RMSD but also the overlap ratio of the side chain contact maps, etc.) and more extensive all-atom optimization protocols (Molecular Dynamics relaxations after the minimization procedures). All these possibilities are now being carefully investigated. The problem of the protein folding "end-game" (model refinement, scoring and selec-

tion) is being intensively pursued in leading laboratories [19], with a promise of success in the near future. Hopefully, the multiscale approach proposed here will contribute to the solution of this problem.

Conclusions

A hierarchical multiscale method of semi-automated modeling of macromolecular assemblies composed of protein and peptides is described and evaluated. The modeling pipe-line starts from Monte Carlo simulation of the assembly process using the CABS reduced model of polypeptides. The crude models are subsequently subject to clustering, all-atom reconstruction and energy minimization with an AMBER 99 force field. It has been shown that the method generates qualitatively correct structures, in many cases of a near-experimental quality. The model selection procedure is capable of picking models that are much better than average from a large set of CABS structures. While acceptable, the model selection is far from being perfect—better models could always be found in the trajectories from the CABS simulations. Thus, the problem of model scoring and selection needs further study. Poor correlation of geometric accuracy with model energy seems be a common deficiency of all reduced-space approaches to protein folding [11, 19].

Finally, it should be mentioned that the performance of the proposed method does not deteriorate significantly when, instead of the experimental structures, the models built by means of comparative modeling are used as the source of the restraints for the receptor. Also, "suspicious" parts of the receptors could be safely made fully flexible. Work in progress is directed towards improvement of the model scoring, and an extension of the method onto non-peptide ligands and onto interactions with nucleic acids.

Acknowledgement This work was partially supported by Polish Ministry of Scientific Research and Information Technology (PBZ-KBN-088/P04/2003).

References

1. Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A (2000) Nat Struct Biol 7:986–990
2. Schneidman-Duhovny D, Nussinov R, Wolfson HJ (2004) Curr Med Chem 11:91–107
3. Gabb HA, Jackson RM, Sternberg MJ (1997) J Mol Biol 272:106–120
4. Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF (2001) Protein Eng 14: 105–113
5. Ewing TJ, Makino S, Skillman AG, Kuntz ID (2001) J Comput Aided Mol Des 15:411–428
6. Claussen H, Buning C, Rarey M, Lengauer T (2001) J Mol Biol 308:377–395
7. Kolinski A (2004) Acta Biochim Pol 51:349–371
8. Swendsen RH, Wang JS (1986) Phys Rev Lett 57:2607–2609
9. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P (2003) J Comput Chem 24:1999–2012
10. Kolinski A, Bujnicki JM (2005) Proteins 61:84–90
11. Vincent JJ, Tai CH, Sathyanarayana BK, Lee B (2005) Proteins 61:67–83
12. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A (2005) Proteins 61:27–45
13. Kmiecik S, Kurcinski M, Rutkowska A, Gront D, Kolinski A (2006) Acta Biochim Pol 53:131–144
14. Kurcinski M, Kolinski A (2006) J Steroid Biochem (in press)
15. Plewczynska D, Kolinski A (2005) Macromol Theor Simul 14:444–451
16. Gront D, Kolinski A (2005) Bioinformatics 21:3179–3180
17. Gront D, Kmiecik S, Kolinski A (2006) J Comput Chem (in press)
18. Claessens M, Van Cutsem E, Lasters I, Wodak S (1989) Protein Eng 2:335–345
19. Bradley P, Misura KM, Baker D (2005) Science 309:1868–1871
20. Vajda S, Camacho CJ (2004) Trends Biotechnol 22:110–116

Praca III (P.III)

Kurciński, Mateusz and Andrzej Koliński. 2010. **Theoretical Study of Molecular Mechanism of Binding TRAP220 Coactivator to Retinoid X Receptor Alpha, Activated by 9-cis Retinoic Acid.** *The Journal of steroid biochemistry and molecular biology* 121(1-2): 124–29.

W pracy zastosowano algorytm giętkiego dokowania białek do badania mechanizmu tworzenia kompleksu molekularnego pomiędzy cząsteczką receptora retinoidów α (RXR α) oraz cząsteczką koaktywatora TRAP220. Przeprowadzono symulacje 5 różnych procesów, które zachodzą lub mogą zachodzić podczas formowania się kompleksu. Na podstawie otrzymanych wyników zaproponowano dwustopniowy mechanizm sekwencyjnego łączenia się receptora z koaktywatorem i cząsteczką kwasu 9-cis retinowego.

Journal of Steroid Biochemistry & Molecular Biology 121 (2010) 124–129

Contents lists available at ScienceDirect

Journal of Steroid Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/jsbmb

Theoretical study of molecular mechanism of binding TRAP220 coactivator to Retinoid X Receptor alpha, activated by 9-cis retinoic acid☆

Mateusz Kurcinski*, Andrzej Kolinski

Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

ARTICLE INFO

Article history:
Received 9 November 2009
Accepted 26 March 2010

Keywords:
Protein interactions
Flexible docking
Molecular modeling
Nuclear receptors

ABSTRACT

Study on molecular mechanism of conformational reorientation of RXR-alpha ligand binding domain is presented. We employed CABS—a reduced model of protein dynamics to model folding pathways of binding 9-cis retinoic acid to apo-RXR molecule and TRAP220 peptide fragment to the holo form. Based on obtained results we also propose a sequential model of RXR activation by 9-cis retinoic acid and TRAP220 coactivator. Methodology presented here may be used for investigation of binding pathways of other NR/hormone/cofactor sets.

1. Introduction

Nuclear hormone receptors (NR) are ligand-activated transcription factors regulating the expression of target genes [1]. They play widespread and important roles in development, metabolism, homeostasis and disease [2,3]. Retinoid X Receptor (RXR) occupies a central position among other nuclear receptors and plays a crucial role as universal heterodimerization partner for many other members of NR superfamily [4]. Its natural ligand—9-cis retinoic acid activates RXR by binding to a pocket located in RXR's ligand binding domain (LBD). Investigation of crystallographic structures of holo and apo forms of RXR-alpha's LBD show significant structural reorientation of the receptor upon ligand binding. Usually nuclear receptors are investigated in respect of their interaction only with primary ligands [5–7], but they also form complexes with other molecules. In holo-form NRs bind with various cell-specific coactivators, which link receptor with the RNA polymerase II—a gear in transcriptional mechanism, while in apo-form NRs form complexes with co-repressors and act as transcriptional suppressors. Abundance of functions of different nuclear receptors and structural similarity between them at the same time makes the NR class very promising pharmacological target.

2. Materials and methods

Two crystallographic structures of RXR-alpha LBD were analyzed. 1lbd is a structure of monomeric apo form [8] and 1xdk holds the structure of holo-RXR-alpha/holo-RAR-beta heterodimer, liganded by 9-cis retinoic acid and bound to 13 amino acid long peptide—TRAP220 coactivator fragment containing LXXLL motif [9]. RXR residues were divided into four separate classes: ligand active site, cofactor active site, structurally conserved residues and others (Fig. 1). Residue was marked as ligand active site if any of its heavy atoms was located within 4.5 Å radius from any of ligand's heavy atoms. Analogues procedure was used to find cofactor active site. Global distance test (GDT) was calculated on 1lbd and 1xdk structures to find fragments that are highly structurally conserved in both structures. GDT cutoff was set to 0.35 Å in order to match the accuracy of casting the structures to the lattice used in CABS model. Residues previously marked as either ligand or cofactor active site were excluded from the conserved residues set. Table 1 presents comparison between these residue classes.

CABS is a reduced model of protein dynamics and thermodynamics, which proved to be extremely effective in various modeling tasks, such as comparative modeling [10,11], protein fragment reconstruction [12], modeling of folding pathways [13–15] and protein docking [16,17]. CABS stands for Cα(CA), Cβ(B) and united pseudoatom located in the center of mass of the side chain (S), which are the only interaction centers. Positions of Cα atoms are restricted to simple cubic lattice, with spacing equal to 0.61 Å, while remaining atoms are located off the lattice. Sampling is controlled by Replica Exchange Monte Carlo (REMC) scheme [18]. Force field was derived from statistical analysis of regularities found in known protein structures. It consists of generic terms

☆ Special issue selected article from the 14th Vitamin D Workshop held at Brugge, Belgium on October 4–8, 2009.

* Corresponding author. Tel.: +48 22 822 02 11x310; fax: +48 22 822 59 96.
E-mail address: mkurc@chem.uw.edu.pl (M. Kurcinski).

0960-0760/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsbmb.2010.03.086

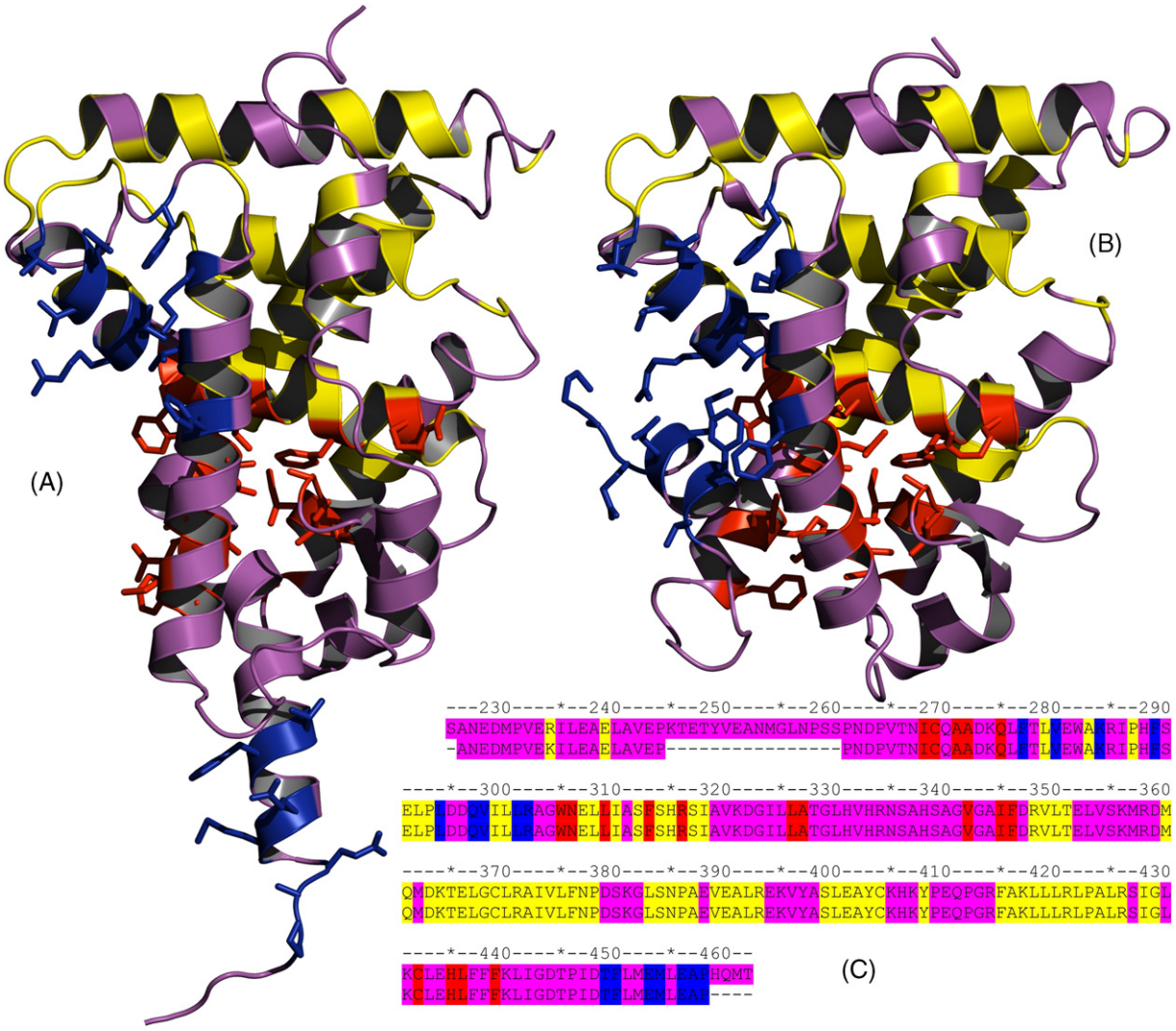


Fig. 1. Crystal structures of RXR-alpha LBD in apo (A) and holo (B) forms. Residues in contact with retinoic acid in apo-RXR are presented in red, those in contact with TRAP220 coactivator in apo-RXR are presented in blue. In yellow most structurally conserved residues (RMSD between apo and holo forms less than 0.35 Å). (C) Alignment of sequences from 1lbd and 1xdk structures, colors as above. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

and biases forcing very flexible Cα-trace to behave in protein-like manner and context-dependant pairwise potentials for side chains' interactions. CABS has been fully automated and integrated with user-friendly interface and data processing/analysis tools [19,20] in modeling platform SPMP, commercially available from Selvita Life Sciences Solutions (http://selvita.com) Command-line version of CABS is available for non-commercial users on our website (http://biocomp.chem.uw.edu.pl). Detailed description of the CABS model was presented earlier [21].

CABS was used to model five molecular transformations which RXR molecule does or hypothetically may undergo. In all five cases receptor molecule consisted of residues 225–462 (238 residues) and in all simulations but first (where only structure of the receptor was modeled) cofactor molecule consisted of residues 641–651 (11 residues). Secondary structure was assigned to all residues in the following way: DSSP [22] was run on both 1lbd and 1xdk structures, whenever secondary structure assignment for 1lbd agreed with the one for 1xdk it was also assigned to the modeled residue, otherwise

Table 1
Structural comparison of 1lbd and 1xdk structures.

Residue selection	Number of residues	RMSD	GDT					
			0.5	1.0	2.0	5.0	7.5	10.0
All	217	7.36	0.57	0.81	0.89	0.97	1.00	1.00
Conserved regions	71	0.34	1.00	1.00	1.00	1.00	1.00	1.00
Ligand active site	19	2.97	0.53	0.74	0.84	1.00	1.00	1.00
Cofactor active site	16	14.79	0.56	0.56	0.56	0.56	0.63	0.75
Other	111	6.28	0.28	0.68	0.83	0.97	1.00	1.00

RMSD (root mean square deviation) is defined as square root of averaged distances between corresponding atoms in two sets. GDT (global distance test) is defined as a number of residues in a subset, for which RMSD is below given threshold, divided by total number of residues.

Table 2
Initial conditions in CABS simulations.

Run	Initial structure		Ligand
	Receptor	Cofactor	
I	apo	–	Present
II	holo	Random	Present
III	apo	Random	Present
IV	apo	Random	Absent
V	holo	Native	Absent

secondary structure was set to coil for that residue. All simulations were run with the same force field parameters set, in the same temperature and for the same number of cycles. Brief description of all five runs is presented below.

2.1. Run I (reorientation of the receptor in the presence of the ligand)

Starting structure of the receptor was taken from 1lbd. Conformational flexibility of the residues marked as conserved (Fig. 1) was strongly restricted by a network of distance restraints imposed on Ca atoms. Distances were measured in both 1lbd and 1xdk structures and subsequently averaged. Similarly, residues marked as ligand active site were restrained as well, but this time only distances from 1xdk were used. This was done to indirectly reflect the presence of the ligand in the binding pocket, since CABS is so far capable of handling only protein molecules. Rest of the residues was left unrestrained.

2.2. Run II (cofactor binding to the receptor–ligand complex)

Starting structure of the receptor was taken from 1xdk. Set of distance restraints imposed on the receptor was the same as in the first run. Initial cofactor conformation was random. Moreover the cofactor molecule was shifted away 15 Å from the receptor surface in respect to the 1xdk structure (all cofactor atoms were shifted by a 15 Å—long vector pointing from receptor's center of gravity (COG) to cofactor's COG). No structural restraints were imposed on cofactor structure.

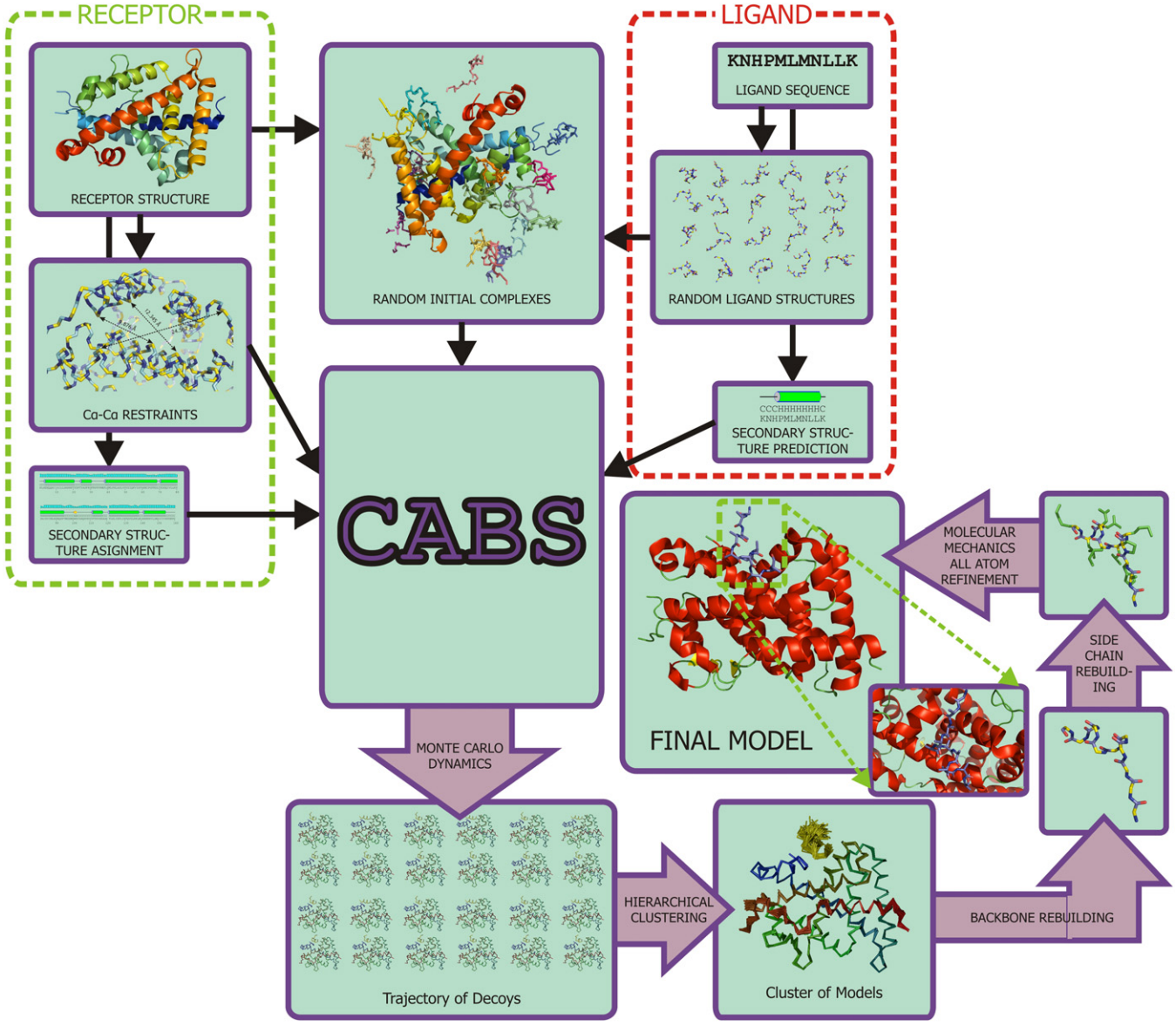


Fig. 2. Modeling scheme used in current work. Complete process is automated and controlled by user-configured scripts.

Table 3
RMSD values calculated on selected models from CABS simulations vs. crystallographic structures of both apo and holo forms of RXR-alpha.

Run	Reference structure	1lbd				1xdk						
	Residue selection	Receptor	Conserved	Ligand active site	Cofactor active site	Receptor	Conserved	Ligand active site	Cofactor active site	Cofactor	Cofactor after receptor superposition	Receptor + cofactor
	Number of residues	217	71	19	16	217	71	19	16	11	11	228
1	Top ranked	7.23	0.50	3.00	14.40	0.79	0.36	0.32	1.97	–	–	–
	Best	5.34	0.43	2.83	10.57	0.58	0.31	0.32	0.96	–	–	–
	Last	7.32	0.50	2.99	14.88	0.78	0.35	0.38	1.83	–	–	–
	Low energy	7.18	0.48	3.04	14.05	0.72	0.35	0.37	1.97	–	–	–
	Average	7.23	0.48	3.01	14.33	0.88	0.35	0.45	2.15	–	–	–
2	Top ranked	7.35	0.81	2.97	14.83	0.64	0.71	0.47	0.70	0.95	1.81	0.74
	Best	4.93	0.42	2.81	10.78	0.58	0.29	0.35	0.28	0.73	1.07	0.66
	Last	7.38	0.80	3.04	14.83	0.64	0.70	0.42	0.63	2.17	2.96	0.91
	Low energy	7.34	0.43	3.00	14.65	0.88	0.31	0.45	1.86	2.75	9.84	2.31
	Average	7.33	0.78	3.01	14.75	0.73	0.69	0.45	0.88	1.62	3.11	0.99
3	Top ranked	7.39	0.48	3.07	14.78	0.80	0.33	0.35	1.41	2.17	10.24	2.38
	Best	3.41	0.43	2.14	6.19	0.76	0.31	0.32	1.19	0.69	4.32	1.95
	Last	7.40	0.48	3.09	14.78	0.81	0.34	0.47	1.41	2.40	10.00	2.34
	Low energy	7.39	0.47	3.09	14.78	0.81	0.34	0.47	1.41	2.40	10.00	2.34
	Average	7.30	0.48	3.02	14.65	0.95	0.35	0.44	1.78	2.57	10.28	2.46
4	Top ranked	0.78	0.33	0.23	1.20	7.25	0.42	2.99	14.57	1.67	16.39	7.80
	Best	0.69	0.33	0.23	0.96	6.22	0.40	2.59	14.24	0.75	4.17	6.20
	Last	0.77	0.33	0.23	1.07	7.24	0.42	2.99	14.56	3.31	13.34	7.51
	Low energy	0.74	0.33	0.23	1.18	7.26	0.42	2.99	14.61	1.77	15.88	7.78
	Average	0.79	0.33	0.24	1.19	7.25	0.42	2.99	14.65	2.86	15.93	7.77
5	Top ranked	5.38	0.32	0.28	10.51	4.76	0.42	2.94	7.32	2.32	14.47	5.48
	Best	4.28	0.30	0.19	6.67	3.04	0.40	2.89	1.88	0.80	2.40	3.19
	Last	7.72	0.30	0.19	15.11	5.20	0.42	2.97	4.16	2.47	10.51	5.61
	Low energy	5.91	0.30	0.19	11.82	5.30	0.42	2.97	6.58	3.60	16.25	6.23
	Average	6.17	0.31	0.26	11.45	5.24	0.42	2.96	6.70	3.24	15.73	6.10

Top ranked, model selected in clustering/refinement procedure; Best, lowest values of listed measures found in all models; Last, last model in CABS trajectory; Low energy, model with the lowest CABS energy; Average, mean value of listed measures, averaged over whole trajectory.

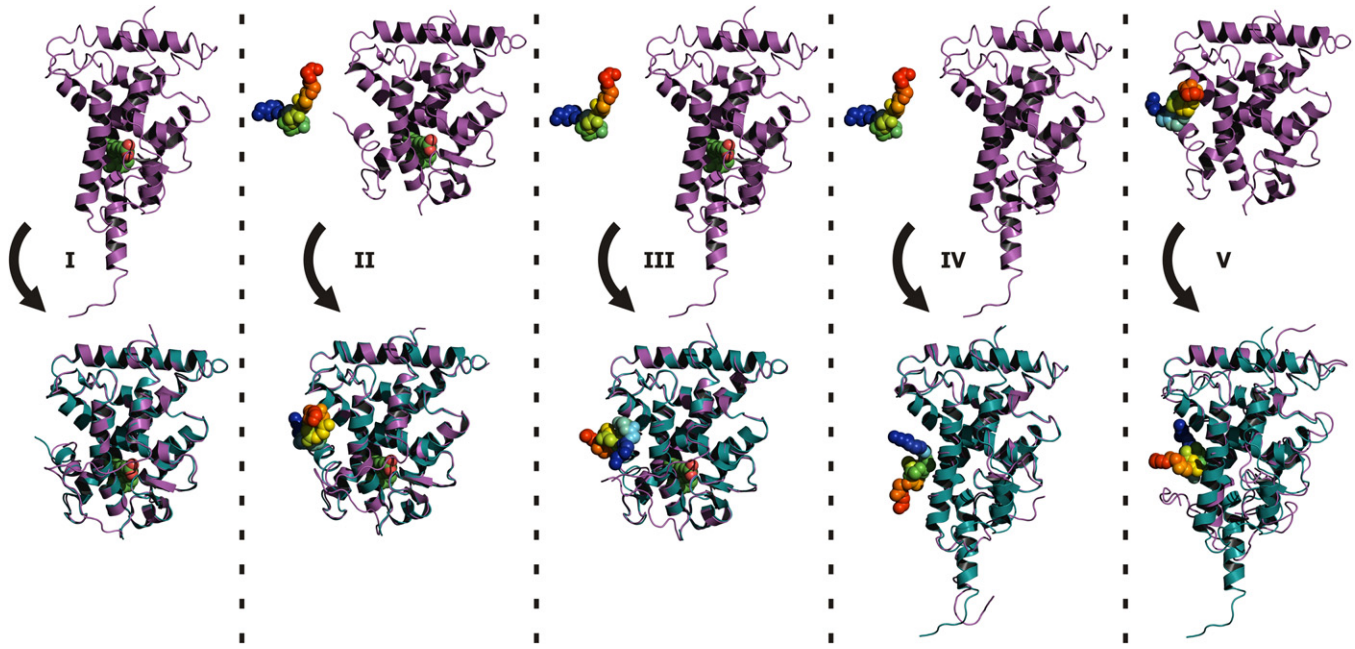


Fig. 3. Initial and final states of the modeled systems in simulations (I–V). In the upper row initial structures of the receptor are shown in purple cartoons. Structure of the cofactor is shown in rainbow spheres. 9-cis retinoic acid—in green. In the bottom row presented are top-ranked models (in purple cartoons (receptor) and rainbow spheres (cofactor)) superposed onto crystallographic structures: 1xdk (I–III) and 1lbd (IV–V). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

2.3. Run III (simultaneous reorientation of the receptor in the presence of the ligand and cofactor binding)

Initial structure of the receptor was the same as in the first run and the cofactor structure was the same as in the second run. Identical set of restraints was used as in the first two runs.

2.4. Run IV (cofactor binding to the receptor in apo form)

Initial structures of both receptor and cofactor were the same as in the third run, but distance restraints imposed on active site residues were derived from 1lbd.

2.5. Run V (holo to apo transformation upon ligand dissociation)

Initial structures of both the receptor and the ligand were taken from 1xdk, but distance restraints imposed on active site residues were derived from 1lbd.

Simulation conditions of all five runs are summarized in Table 2. Every run consisted of ten separate iterations differing only by random seed. Every iteration produced 1000 models to sum up to total 10 000 models per run. These models were grouped together by their mutual similarity in two-stage hierarchical clustering procedure [23]—first within 1000 models produced in individual iterations, next medoids of the clusters found in the first stage were clustered again. Medoids of top three biggest clusters were further refined: first reconstructed to all-atom representation using BBQ [24] and SCWRL3.0 [25] and energy-minimized in empirical force field Amber99 [26]. Final model for each run was selected according to the final energy after the minimization [27]. Scheme of the described methodology is presented in Fig. 2.

3. Results and discussion

3.1. Run 1 (reorientation of the receptor in the presence of the ligand)

The process of receptor reorientation was modeled with high accuracy, as confirmed by RMSD between top-ranked model and 1xdk structure in all categories. Cofactor binding site was modeled with 1.97 Å accuracy in respect to the 1xdk structure, which suggests, that additional to ligand-induced conformational adjustment is required upon binding with the cofactor.

3.2. Run 2 (cofactor binding to the receptor–ligand complex)

RMSD between top-ranked model and 1xdk structure calculated on all residues is below 0.8 Å. Additional adjustment in the cofactor binding site is reflected in RMSD decrease to 0.70 Å in that region, when compared to the model obtained in the first run. This suggests that hormone and cofactor activate the receptor sequentially rather than simultaneously.

3.3. Run 3 (simultaneous reorientation of the receptor in the presence of the ligand and cofactor binding)

Although generally low RMSD on the complete structure (2.38 Å), in this run top-ranked model had incorrect orientation of the docked cofactor (RMSD on the cofactor after superposition of the receptor with 1xdk structure was over 10 Å). Furthermore, cofactor itself was in wrong conformation (RMSD 2.17 Å in respect to the 1xdk). The shape of the cofactor binding site was deformed as well (RMSD 1.41 Å). This is another premise suggesting sequential binding of the ligand/cofactor molecules.

3.4. Run 4 (cofactor binding to the receptor in apo form)

In this case receptor remained in its apo form, which inactivated it for cofactor reception as shown by high RMSD values.

3.5. Run 5 (holo to apo transformation upon ligand dissociation)

This case was run to model holo to apo transformation upon ligand dissociation. Indeed shape of both ligand and cofactor active sites has changed significantly. Also cofactor structure and location have been changed. However, apo form was not reconstructed completely—final model shows some structural similarities to both apo and holo forms (RMSD 5.38 Å and 4.76 Å respectively).

In runs I and II the process of binding the ligand and subsequently the cofactor was modeled with great accuracy. At the same time runs III and IV demonstrate that different sequence of binding leads to wrong conformations of final structures. RMSD values are presented in Table 3. Initial and final structures from all simulations are shown in Fig. 3.

4. Conclusions

We present a methodology for investigation of receptor–cofactor binding mechanisms. Case study included prediction of binding TRAP220 coactivator to RXR-alpha receptor in both holo and apo forms, in simultaneous with the ligand, or sequential order. Also ligand-activated apo–holo transformation was investigated. Obtained results remain in agreement with experimental data (crystallographic structures), therefore similar procedure may be applied to investigation of binding pathways in other NR/hormone/cofactor complexes. We also propose a two-stage sequential mechanism of RXR-alpha activation by 9-cis retinoic acid and TRAP220 coactivator, as this matter still remains unresolved. We hope that methodology presented here may find wide spectrum of applications as it provides fast and inexpensive alternative to experimental techniques.

Acknowledgements

This work was supported by the NIH grant no. 1R01GM081680 and Polish Ministry of Science and Higher Education, grant no. NN301465634. Computational part of this work was done using the computer cluster at the Computing Center of Faculty of Chemistry, University of Warsaw. A commercial version of CABS-based modeling software was used (http://www.selvita.com/selvita-protein-modeling-platform.html).

References

[1] R.M. Evans, The steroid and thyroid hormone receptor superfamily, Science 240 (4854) (1988) 889–895.
[2] D.J. Mangelsdorf, et al., The nuclear receptor superfamily: the second decade, Cell 83 (6) (1995) 835–839.
[3] A. Chawla, et al., Nuclear receptors and lipid physiology: opening the X-files, Science 294 (5548) (2001) 1866–1870.
[4] D.J. Mangelsdorf, R.M. Evans, The RXR heterodimers and orphan receptors, Cell 83 (6) (1995) 841–850.
[5] P. Rotkiewicz, et al., Model of three-dimensional structure of vitamin D receptor and its binding mechanism with 1alpha,25-dihydroxyvitamin D(3), Proteins 44 (3) (2001) 188–199.
[6] W. Sicinska, P. Rotkiewicz, H.F. DeLuca, Model of three-dimensional structure of VDR bound with Vitamin D3 analogs substituted at carbon-2, J. Steroid Biochem. Mol. Biol. 89–90 (1–5) (2004) 107–110.
[7] R.R. Sicinski, et al., 2-Ethyl and 2-ethylidene analogues of 1alpha,25-dihydroxy-19-norvitamin D(3): synthesis, conformational analysis, biological activities, and docking to the modeled rVDR ligand binding domain, J. Med. Chem. 45 (16) (2002) 3366–3380.
[8] W. Bourguet, et al., Crystal structure of the ligand-binding domain of the human nuclear receptor RXR-alpha, Nature 375 (6530) (1995) 377–382.
[9] V. Pogenberg, et al., Characterization of the interaction between retinoic acid receptor/retinoid X receptor (RAR/RXR) heterodimers and transcriptional coactivators through structural and fluorescence anisotropy studies, J. Biol. Chem. 280 (2) (2005) 1625–1633.
[10] A. Kolinski, J.M. Bujnicki, Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models, Proteins 61 (Suppl. 7) (2005) 84–90.
[11] A. Kolinski, D. Gront, Comparative modeling without implicit sequence alignments, Bioinformatics 23 (19) (2007) 2522–2527.
[12] M. Boniecki, et al., Protein fragment reconstruction using various modeling techniques, J. Comput. Aided Mol. Des. 17 (11) (2003) 725–738.
[13] S. Kmiecik, A. Kolinski, Characterization of protein-folding pathways by reduced-space modeling, Proc. Natl. Acad. Sci. U.S.A. 104 (30) (2007) 12330–12335.
[14] S. Kmiecik, et al., Denatured proteins and early folding intermediates simulated in a reduced conformational space, Acta Biochim. Pol. 53 (1) (2006) 131–144.
[15] S. Kmiecik, A. Kolinski, Folding pathway of the b1 domain of protein G explored by multiscale modeling, Biophys. J. 94 (3) (2008) 726–736.
[16] M. Kurcinski, A. Kolinski, Hierarchical modeling of protein interactions, J. Mol. Model. 13 (6–7) (2007) 691–698.
[17] M. Kurcinski, A. Kolinski, Steps towards flexible docking: modeling of three-dimensional structures of the nuclear receptors bound with peptide ligands mimicking co-activators' sequences, J. Steroid Biochem. Mol. Biol. 103 (3–5) (2007) 357–360.
[18] R.H. Swendsen, J.S. Wang, Replica Monte Carlo simulation of spin glasses, Phys. Rev. Lett. 57 (21) (1986) 2607–2609.
[19] D. Gront, A. Kolinski, Utility library for structural bioinformatics, Bioinformatics 24 (4) (2008) 584–585.
[20] D. Gront, A. Kolinski, BioShell—a package of tools for structural biology computations, Bioinformatics 22 (5) (2006) 621–622.
[21] A. Kolinski, Protein modeling and structure prediction with a reduced representation, Acta Biochim. Pol. 51 (2) (2004) 349–371.
[22] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637.
[23] D. Gront, A. Kolinski, HCPM—program for hierarchical clustering of protein models, Bioinformatics 21 (14) (2005) 3179–3180.
[24] D. Gront, S. Kmiecik, A. Kolinski, Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates, J. Comput. Chem. 28 (9) (2007) 1593–1597.
[25] M.J. Bower, F.E. Cohen, R.L. Dunbrack Jr., Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool, J. Mol. Biol. 267 (5) (1997) 1268–1282.
[26] E.J. Sorin, V.S. Pande, Exploring the helix-coil transition via all-atom equilibrium ensemble simulations, Biophys. J. 88 (4) (2005) 2472–2493.
[27] S. Kmiecik, D. Gront, A. Kolinski, Towards the high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field, BMC Struct. Biol. 7 (2007), p. 43.

Praca IV (P.IV)

Kurciński Mateusz, Andrzej Koliński and Sebastian Kmiecik, 2013. **Mechanism of folding and binding of an intrinsically disordered protein derived from unbiased simulations.** (zgłoszona do druku)

Celem pracy było opisanie mechanizmu, według którego białko o nieuporządkowanej strukturze zwija się do swojej aktywnej konformacji podczas tworzenia kompleksu z innym białkiem. Jako układ modelowy wybrano szeroko opisany w literaturze kompleks KIX/pKID.

W celu zbadania mechanizmu przeprowadzono symulacje giętkiego dokowania pKID do KIX za pomocą modelu CABS. Analiza otrzymanych wyników wskazuje na mechanizm, w którym najpierw pKID łączy się niespecyficznie z receptorem poprzez liczne nienatywne kontakty hydrofobowe, a dopiero po utworzeniu kluczowych połączeń natywnych następuje specyficzne wiązanie z jednoczesnym przyjęciem przez pKID natywnej konformacji. Mechanizm ten przypomina klasyczny mechanizm nukleacji-kondensacji, według którego zwija się większość białek globularnych.

Mechanism of Folding and Binding of an Intrinsically Disordered Protein as Revealed by Ab Initio Simulations

Mateusz Kurcinski[1], Andrzej Kolinski[1] and Sebastian Kmiecik[1,*]

[1] Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

* To whom correspondence should be addressed. Tel: +48 22 822 02 11 ext. 310; Fax: +48 22 822 02 11 ext. 320; Email: sekmi@chem.uw.edu.pl

Abstract

A complex of the phosphorylated kinase-inducible domain (pKID) with its interacting domain (KIX) is a model system for studies of mechanisms by which intrinsically unfolded proteins perform their functions. These mechanisms are not fully understood. Using an efficient coarse-grained model, ab initio simulations were performed of the coupled folding and binding of the pKID to the KIX. The simulations start from an unbound, randomly positioned and disordered pKID structure. During the simulations the pKID chain and its position remain completely unrestricted, while the KIX backbone is limited to near-native fluctuations. Ab initio simulations of such large-scale conformational transitions, unaffected by any knowledge about the bound pKID structure, remain inaccessible to classical simulations. Our simulations recover an ensemble of transient encounter complexes in good agreement with experimental results. We find that a key folding and binding step is linked to the formation of weak native interactions between a preformed native-like fragment of a pKID helix and KIX surface. Once that nucleus forms, the pKID chain may condense from largely disordered encounter ensemble to a natively bound and ordered conformation. The observed mechanism is reminiscent of a nucleation-condensation model, a common scenario for folding of globular proteins.

Introduction

Proteins are dynamic molecules and may sample a huge ensemble of sometimes vastly different conformations, while participating in molecular recognition events. Over the last fifty years, the understanding of molecular recognition processes has been dominated by the 'induced fit' hypothesis [1]. According to this hypothesis it is the binding interaction that drives a protein from unbounded to bound conformation. In contrast, a number of recent experimental studies suggest a 'conformational selection' mechanism postulating that all protein conformations preexist in the ensemble sampled by the free protein in the absence of the binding partner [2]. Noteworthy, binding mechanisms can be very complex and the distinction between 'conformational selection' and 'induced fit' may be blurred [3]. For instance, the mechanistic path may depend on many factors (e.g. the mechanism of ligand binding changes with varied ligand and protein concentrations [3]) and thus can be accessed using detailed knowledge of the kinetics and thermodynamics of the reaction mechanism.

Many regulatory proteins are conformationally disordered under physiological conditions and form ordered structures only upon target binding [4-6]. Intrinsically disordered proteins (IDPs) remain difficult to study, both experimentally [7] and by simulation [8], due to their high heterogeneity. The recent review of known disorder-based complexes suggests that the general model for the interaction of IDPs with their partners should be considered as a two stage process [9], with the first stage leading to the formation of disordered complexes and a slower second stage resulting in the formation of ordered associations. Based on theoretical analysis, it has been proposed that folding and binding processes may be enhanced by the 'fly casting' mechanism [10]. In this scenario, the disordered protein binds weakly at a relatively large distance to its target and folds while approaching the binding site. It has been hypothesized that an disordered protein can have a greater capture radius for a native binding site; therefore the binding rate can be significantly enhanced in comparison with the fully folded protein [10].

Molecular simulations of IDPs are challenging or impossible to perform using contemporary simulation techniques, especially if they are carried in an ab initio fashion (no experimental information about the simulated system is used). The challenge can be characterized as efficient treatment of substantial conformational changes combined with a sufficiently accurate model to capture the relevant properties of the system [8]. This problem becomes even more complex when IDP binding simulation is sought. As highlighted in the recent reviews of the results of the Critical Assessment of Predicted Interactions (CAPRI) [11, 12], a community-wide experiment aimed to assess the state of the art in protein-protein

docking, the major unsolved problem in the field is the docking of proteins with a substantial backbone conformation change. These challenges can be addressed using efficient sampling strategies combined with coarse-grained models [13].

Most (or perhaps all) of the published simulations of coupled folding and binding of pKID-KIX used simulation models in which pKID was biased towards the native-like state. This bias is usually introduced either by the use of natively-biased force-field [14-19] or by sampling limited to native complex arrangement [20, 21]. In this work, we describe simulation results obtained in an ab initio (unbiased) fashion using an efficient coarse-grained simulation model – the CABS model. Based on the generated conformational ensembles we describe a complete mechanism of successful binding, starting from a barely bound and highly disordered ensemble, and involving a multitude of non-native interactions on the path to native binding.

Materials and Methods

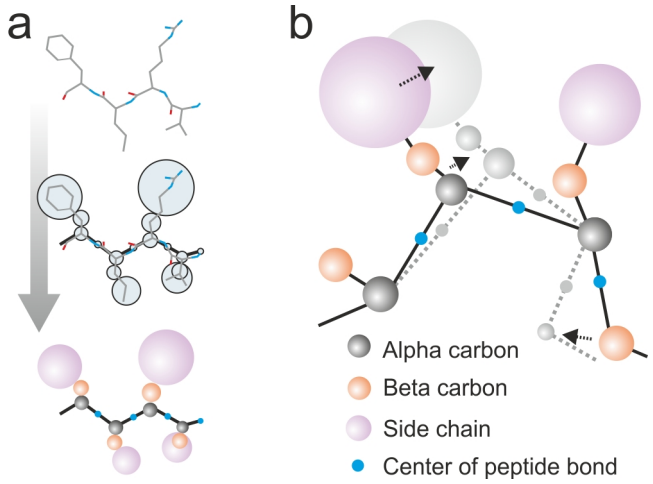
Simulation model

CABS is a coarse-grained protein model, extensively tested in protein structure prediction [22-24] and simulations of protein dynamics: folding mechanisms [25-29] and near-native fluctuations [30, 31]. CABS has also been used for flexible protein-protein docking [32, 33] and, in conjunction with other computational tools and experimental data, for the assembly of the first realistic model of the human telomerase enzyme [34]. The flexible docking tests included a study of the binding mechanism of the peptide TRAP220 co-activator to the Retinoid X Receptor [33] in which the modeled system was treated in a similar manner as pKID-KIX in the present work. Namely, during the docking simulation, the peptide co-activator chain (11 amino acids in length) was treated as fully flexible and allowed for unbiased detection for the binding site, while the receptor structure (238 amino acids in length) was restricted to near-native fluctuations (see Movie S1 in the Supporting Material showing example simulation trajectory). The resulting mechanism agreed well with the experimental data.

As CABS has been described in great detail before [35], here we only outline its main features. CABS uses reduced representation of protein chains limited to up to four atoms per residue (see Figure 1): Cα carbon, Cβ carbon (except for Glycine), a united pseudo-atom representing the amino acid side chain (except for Glycine and Alanine) and a pseudo atom representing the center of a peptide bond. In addition, CABS employs discreet space representation, by limiting Cα atom positions only to cubic lattice. The distance between lattice beads is arbitrarily set to 0.61 Å, which is small enough to prevent any lattice-

associated inaccuracies, but large enough to allow for a significant boost in calculations at the same time. Importantly, the resolution of CABS-generated models enables reconstruction to realistic all-atom models [28, 36]. The CABS force field is purely statistical, derived from a representative set of known protein structures. Conformational sampling proceeds by small, local transitions controlled by the very efficient Replica Exchange Monte Carlo algorithm. As demonstrated in the previous reports, the long series of such moves may faithfully reproduce the realistic protein dynamics of large time-scale processes [25-28]. Energy minimization is handled by simulated temperature annealing. The degree of protein flexibility can be adjusted in CABS from almost completely rigid to fully unrestrained, both locally and globally. In particular, it allows docking a fully flexible protein/peptide ligand to the receptor fluctuating around near-native conformations. All these enhancements are aimed at modeling bigger systems in longer time-scales without compromising accuracy.

Figure 1. Overview of the simulation model (CABS). (a) coarse-grained representation, (b) example move.



Simulations setup

In our simulations, we employed an efficient Replica Exchange version of the CABS Monte Carlo sampling scheme with simulated temperature annealing. In a single simulation, we used twenty replicas uniformly distributed on the temperature scale, with all replicas starting from random conformations. To ensure exhaustive sampling of the conformational space, we ran 75 independent simulations. In these simulations, different temperature ranges were applied in search of optimal conditions at which folding and binding occurs. For a single simulation, an initial setup was constructed in the following manner:

(i) KIX structure was taken from model 1 from the 1KDX pdb entry.

(ii) pKID structure was randomly generated and randomly placed on the surface of a 50Å radius sphere centered at the KIX center of gravity. No restraints were imposed on the pKID molecule, leaving it fully flexible. Weak distance restrains were imposed on the KIX structure, thus keeping it in near-native conformation, but also allowing some degree of flexibility (note that CABS dynamics was reported to be a good predictor of near-native flexibility [30, 31]). Therefore, no information about the pKID molecule's native conformation or its binding location was used.

(iii) For simplicity, CABS was designed to handle protein molecules with the 20 standard amino acids only. To simulate the effect of phosphorylated Ser133 in the pKID molecule, we substituted serine in position 133 with glutamic acid, which is a common approach for mimicking serine phosphorylation.

A single simulation of twenty replicas took around fifteen hours on a single CPU.

Simulation pre-analysis and selection

The structural analysis of the resulting trajectories showed that pKID binds to multiple sites of the KIX surface, which is in accordance with competition and mutagenesis approaches combined with NMR titrations and ^{15}N relaxation dispersion experiments [37]. According to these experimental data, pKID binds natively and non-natively to additional hydrophobic sites on the KIX surface, including the site for the other KIX-interaction domain (located at the opposite KIX surface to the pKID site).

Furthermore, we analyzed energetic properties of all the resulting complexes using CABS energy values for: the pKID-KIX complex, or pKID alone, or pKID-KIX interaction. We found that using CABS energy values it is not possible to discriminate pKID-KIX complexes, which are closest to the native. Generally speaking, some of the non-native complexes were scored on the same level as those closest to the native. According to experimental data [37], the native interaction is of high affinity relative to other non-native sites being of low affinity. Interestingly, the effect of high affinity binding comes from the fine details of pKID-KIX interactions. Namely, it is due to intermolecular interactions of the phosphate moiety attached to Ser133, since non-phosphorylated KID binds natively to KIX, but with low affinity [38]. Therefore, the native pKID-KIX structure seem correspond to the

local energy minimum, which is flattened in CABS energy landscape (also note that phosphorylated Ser133 is mimicked in our simulations by glutamic acid).

Since our goal was to get an insight into the mechanism of native pKID-KIX binding, for further analysis we attempted to select the simulations, which ended up in a near-native pKID-KIX arrangement. The selection was based on ranking using the highest fraction of near-native complex structures (i.e. RMSD below 5 Å vs. model 1 from the 1KDX pdb entry) among 10% of the terminal trajectory frames. Based on these criteria, ten trajectories were chosen, each one of 1000 snapshots and having at least 50 near-native structures in the final hundred (all the snapshots of the selected trajectories are characterized in the Figure 2). Noteworthy, using replica exchange sampling, it is not possible to identify the correct folding and binding mechanisms from a replica trajectory. This is because a single replica performs a random walk in temperature space. However, if temperature changes are small over the observed folding and binding events, as in our simulations, the replica trajectories can be considered as correct approximations of the system dynamics.

RMSD calculations

In this paper, we used RMSD (Root Mean Square Deviation) calculated on Cα atoms of the entire pKID-KIX complex as a measure of structural similarity between simulated pKID-KIX complexes and the native complex. Since the KIX molecule was restricted only to small, near-native fluctuations, the KIX structure variations did not affect the RMSD values.

Results

pKID-KIX complex – a model system for IDP binding studies

To investigate the mechanism of coupled IDP folding and binding, we used a protein docking methodology utilizing the CABS protein dynamics model [25] (see Methods). We used this approach to simulate the binding of the phosphorylated kinase-inducible domain (pKID) of the transcription factor CREB to the KIX domain of the CREB binding protein.

The pKID domain is a 28 residue peptide which folds upon binding to the KIX domain [37, 39, 40]. NMR spectra of pKID in an unbounded form have the characteristics of an disordered peptide with a slight propensity towards helix formation in the KIX binding region [39]. In the pKID-KIX complex (pdb code: 1KDX), pKID is folded into two helices designated as αA and αB (residue number 120 to 129 and 133 to 145, respectively) [39]. These two helices are arranged in the complex structure at an angle of about 90 degrees and are essentially wrapped

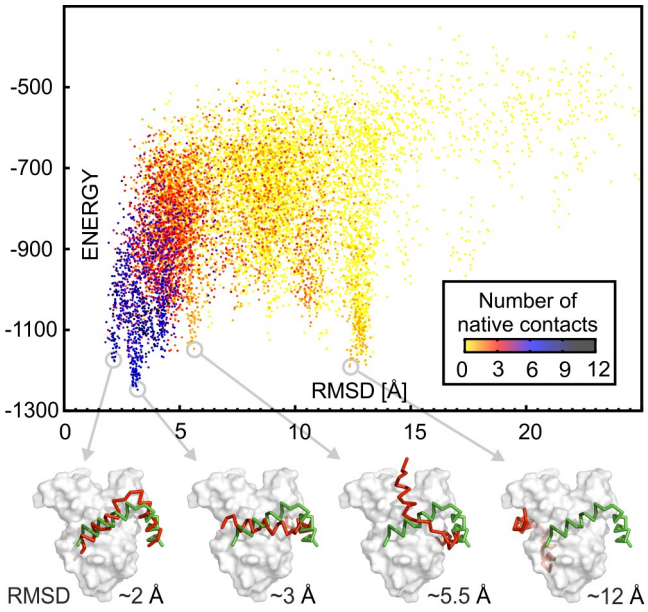
around the C-terminal alpha helix (α3) of KIX. KIX is composed of three alpha helices (from the N- to C- terminus: α1, α2 and α3). KIX α1 and α3 helices form together a hydrophobic binding patch for the pKID αB helix, while the second pKID alpha helix (αA) interacts with a different side of the KIX α3 helix.

The mechanism of pKID and KIX binding has been a subject of numerous simulation studies. As highlighted in the introduction, the major obstacle to simulating the IDP binding process is efficient treatment of large timescale dynamics, while preserving sufficient model accuracy. The gold standard of accuracy for protein simulations is all-atom explicit solvent molecular dynamics (MD), recently used for the characterization of pKID-KIX [20]. Due to timescale limitations, the MD simulations of pKID-KIX binding were carried out as high-temperature unfolding: thus, an assumption was necessary that the mechanisms of high-temperature unfolding and room-temperature folding are similar. The problem of sampling efficiency may be overcome by the application of Go-modeling as demonstrated in pKID-KIX studies [14-18]. The Go-like models provided valuable insights into characteristics of IDPs, e.g. into advantages of IDPs in molecular recognition through the 'fly-casting' mechanism [10]. The major assumption under the Go model is that interactions found in the native state also prevail in the studied mechanisms. However, the effect of non-native interactions on protein folding has been demonstrated in many experiments, and is expected to be even more pronounced for IDPs than for globular proteins. This has been demonstrated by studies using a simple Go-like model with a non-native hydrophobic interaction component [16, 17], which suggested that non-native interactions might accelerate the binding rate.

Encounter complexes in pKID and KIX binding

To enable thorough understanding of the mechanisms of native binding, we jointly analyzed 10 representative simulation trajectories, which ended up in near-native association (for details, see Methods). These simulations are characterized in Figure 2. As presented in the figure, the trajectories show multiple minima (representing near- native associations or a non-native complex on the opposite surface of the pKID binding site) and a multitude of medium- and high- energy transient states. These kinds of transient states are characteristic for diffusional encounter complexes [41], an ensemble of molecular arrangements at the end point of diffusional association which involve short-range interactions and non-diffusional motion in search of favorable orientations and contacts.

Figure 2. Characteristics of the simulation trajectories. Each point in the plots represents: pKID-KIX energy value (vertical axis), RMSD value (root mean-square deviation to the native pKID-KIX complex; horizontal axis), colored according to the number of native contacts between pKID and KIX. Additionally, example low energy KIX-pKID models are shown (pKID chain is shown in red, together with its native conformation in green).



In Figure 3a, we further analyze the simulation trajectories using the distribution of RMSD values. The histogram analysis shown allowed us to define the following ensembles of complexes (reflected in three significant clusters of peaks):

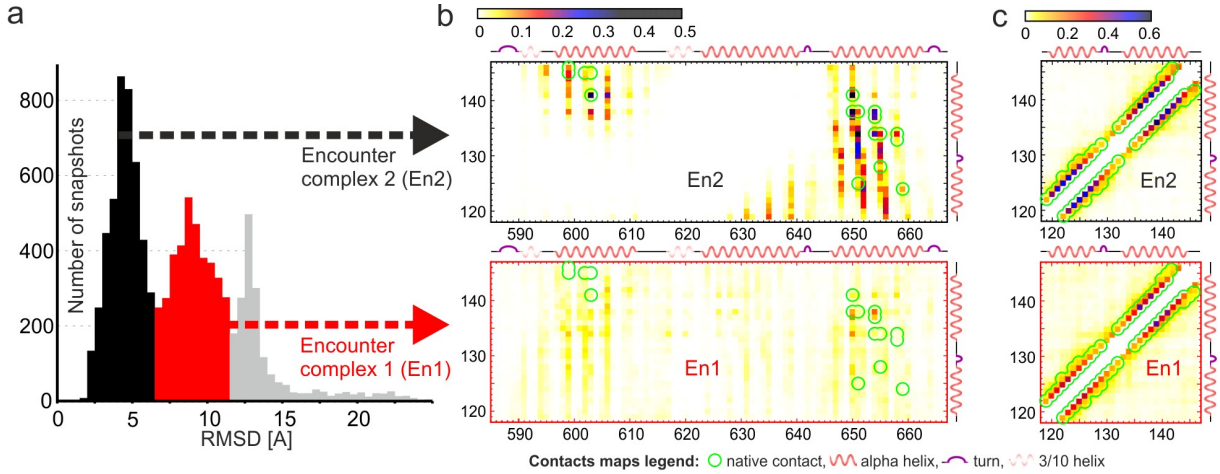
(i) encounter complexes 1 (En1) – representing non-native or partially-native complexes, assembled in the neighborhood of the native site (with RMSD values in the range of 6.5Å to 11.5 Å, constituting 37% of the trajectories)

(ii) encounter complexes 2 (En2) – representing native or partially-native complexes, assembled at least partially in the native site (with RMSD values lower than 6.5Å, constituting 44% of the trajectories)

(iii) unbounded/non-native complexes representing unbounded or partially-native complexes, assembled unlike the pKID native orientation (with RMSD values larger than 11.5Å, constituting 19% of the trajectories)

Below, we will focus on the analysis of encounter complexes En1 and En2, the two dominating ensembles, which comprise transition paths between the very first stage of binding and near-native pKID and KIX arrangements.

Figure 3. Characteristics of dominant structural ensembles by residue-residue contact maps. (a) distribution of RMSD values of the pKID-KIX complex (root mean-square deviation to the native pKID-KIX complex). The distribution suggests the presence of three ensembles marked as: Encounter 1 (En1, bars colored in black), Encounter 2 (En2, red) and unbounded/non-native complexes (gray). (b) Residue contact maps for pKID and KIX interactions. (c) Residue contact maps for intra-pKID interactions. Contact maps in (b) and (c) show contact frequencies for En2 (upper row) and En1 (lower row) complexes. Contact frequencies are denoted by colors (see color legends). Residue numbers and secondary structure are marked on map borders. Native contacts are outlined in green circles. In (c) maps, short-range contacts (up to *i*, *i*+2) are omitted for clarity. Contact maps were derived from distances between the gravity centers of the side chains using 5Å cut-off.



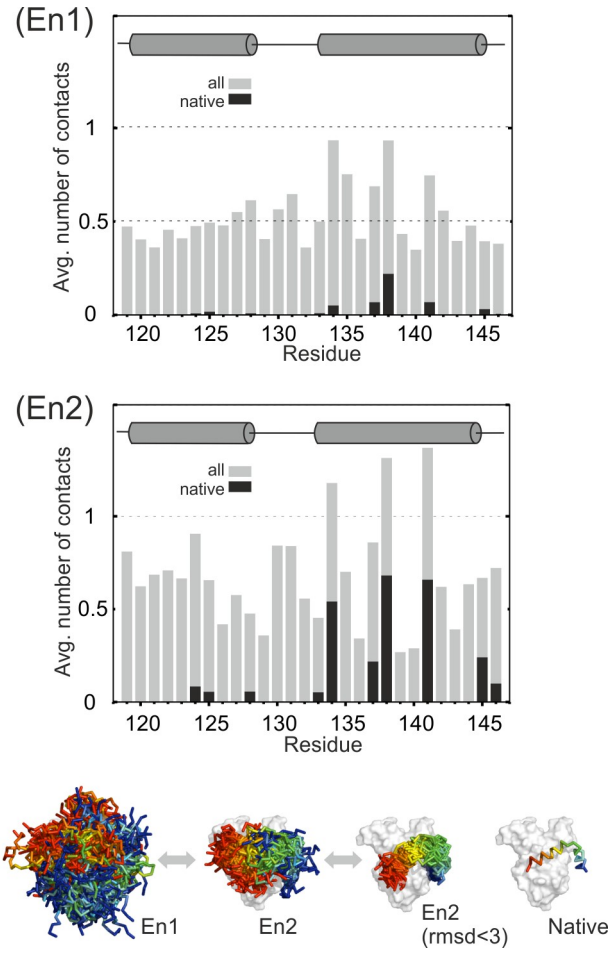
Characteristics of encounter complex 1 (En1)

In Figure 3, we present a detailed insight into the complexity of En1 and En2 ensembles, on the level of residue-residue contacts. The figure confronts En1 and En2 contact maps calculated for pKID and KIX interactions (b) and intra-pKID interactions (c). In En1, the average contact frequency between any pKID and KIX residue is at a very low level (Figure 3b, bottom map). This is indicated by a few contacts marked in red/orange (contact frequency around 0.1) and many contacts colored in yellow (contact frequency around 0.05). As shown in the map, En1 is clearly dominated by a multitude of non-native contacts. Interestingly, these contacts are mostly formed by hydrophobic KIX residues, which are involved in native binding (see vertical yellow lines in the map, in KIX 598-608 and 647-662 regions, and native contact positions marked by green circles). Except for non-native binding, a tendency to form some of the native contacts can already be observed. In particular, two native contacts of Leu138 with Ala654, and with Tyr650, seem to form an anchoring spot for further binding, as these belong to the most frequent contacts in the En1 set (Leu138-Ala654

being the most frequent; the list of the most frequent contacts in En1 is presented in Table S1 in the Supporting Material). In addition to interacting with KIX, Leu138 is engaged in the formation of the most persistent fragment of pKID secondary structure in the En1 ensemble: the αB helix fragment between 136 and 141 residue. This is indicated by the most frequent intra-chain contacts of pKID (Lys136-Asn139 and Leu138-Leu141) present in about 40% En1 snapshots (see Figure 3c, bottom panel). Other contacts (i, i+3, reflecting the formation of αA and αB helices) have contact frequency in the range of 0.3-0.1. The magnitude and pattern of intra-pKID contacts suggest that the pKID is generally disordered, and that the αA and the αB regions are partly helical, with the αB region being more helical than αA.

The important role of Leu138 is also revealed in the analysis of the involvement of pKID residues in En1 complex formation, see Figure 4. As shown in the figure, Leu138 forms native contacts with KIX significantly more frequently than any other pKID residue. The pattern of the average number of non-native contacts per residue seems to reflect the disordered and highly flexible nature of pKID, since it is uniformly distributed around 0.4 among all of the residues. The only exception is Tyr134, Leu138 and Leu141 (together with their neighboring residues) that make a noticeably larger number of contacts, with a small fraction being native.

Figure 4. pKID involvement in complex assembly. The plots present an average number of all contacts (gray bars) and native contacts (black bars) for pKID residues upon binding to KIX. Two plots are shown, exclusively for the encounter complexes: En1 (upper plot) and En2 (lower plot). At the bottom, structural visualization of the following ensembles is presented: En1, En2, near-native from En2 (RMSD<3) and native pKID-KIX complex structure.



Characteristics of encounter complex 2 (En2)

The contact map characterization of the En2 complex shows two patches of contacts located just where the binding site is (Figure 3b, top map). As presented in the map, most of the native contacts occur with the highest observed frequencies. At the same time the En2 complex is stabilized by frequently occurring non-native contacts. The moderate frequency of the most pronounced contacts (in a range of 0.35-0.2) and abundance of various non-native interactions both demonstrate the transient character of the complex being formed. Native contacts identified as anchoring in En1 still belong to the most pronounced ones; however, in En2, pKID Leu141 seems to have the major anchoring role. Namely, two native contacts of Leu141 belong to the four of the most frequent contacts in En2: Tyr134-His651, Leu141-

Tyr650, Leu138-Tyr650 and Leu141-Leu603 (the list of the top frequent contacts in En2 is presented in Table S1 in the Supporting Material). Thus, Leu141 stabilizes the interaction of Leu138 with Tyr650 (already pronounced in En1) and simultaneously anchors the pKID to another helix in KIX by contacts with Leu603. These interactions seem to be supported by other native hydrophobic contacts made by Tyr134, Ile137 and Leu138 with KIX Ala654. As demonstrated in Figure 3c, the pKID chain is significantly more helical in En2 than in the En1 complex. This folding upon binding effect is also demonstrated by the absence of cross-helix contacts (present in En1). The essential role of the hydrophobic residues: Tyr134, 138Leu and Leu141 in native binding is also highlighted in Figure 3 (lower panel). These residues located on one face of the αB helix come into contact with KIX most frequently. This is largely due to their native contacts, with the number of exclusively native contacts much higher than for any other pKID residue (Figure 4).

In the bottom of Figure 4, we compared structural visualizations of the En1 and En2 complexes found in our simulations, together with the native complex structure. In addition, an ensemble of En2 complexes with RMSD lower than 3 Angstroms is also presented. In these nearest to the native complexes, the orientation of the αA helix is less defined than that of αB. This is consistent with the NMR characterization of the pKID/KIX complex structure [39] in which αB helix interactions and orientation relative to KIX are very well defined, while those of αA are not (note that, consequently, the definition of αA native contacts may be not so accurate as that of αB).

In accordance with the above facts, native contacts of pKID αA with KIX are the least pronounced on the maps of native contact occurrence during example simulations (see Figure S1 in the Supporting Material). As shown in Figure S1, transitions from En1 to En2 complexes (reflected in the shift of RMSD values from around 10 to 5 Angstroms) are involved in the cooperative formation of most of the native interactions. This contrasts with En1 complexes, which interact simultaneously usually with one or two clusters of native contacts (formed by residues 124-128 in αA; or by residues 133-138 or 141 or 145-146 in αB).

Discussion

In agreement with NMR observations [37], our simulations indicate that pKID forms an ensemble of transient encounter complexes upon native binding to the KIX domain. We show

that the generated encounter complexes may be separated into two structurally different ensembles, En1 and En2: the first being distant from the native pKID arrangement, and the second more native-like (which comprises ordered, near-native complexes, but is mostly disordered-like). The transition between the En1 and the En2 ensembles reflects a key step for the native association. Most of the native hydrophobic interactions in the binding interface form later, and rather cooperatively, after the key transition event (see Figure S1 in the Supporting Material). Such a binding pathway in which native interactions form late is consistent with the previous studies on the pKID/KIX binding mechanism using NMR [37] or a topology-based Go-model [17], and also with studies of the binding of other IDPs using stopped-flow spectroscopy [42-44], or phi value analysis [45].

More importantly, the picture seen in our simulations agrees very well with the experimental identification of pKID residues that are essential for the binding. All of the pKID residues that we found responsible for making initial native contacts (lying in the region of the αB helix: Tyr134, Leu138, Ile137 and Leu141, see Figure 4) have been found in an NMR relaxation-dispersion experiment as critical for productive binding [37]. Moreover, the NMR data suggested that the αB helix region is partly helical (up to 30%) in the low-affinity complex (counterpart of the En1 ensemble in our simulations), which agrees quantitatively with our findings. An important role of the above-mentioned hydrophobic residues has also been demonstrated in mutation analyses [38, 40, 46]. For instance, Ile137, Leu138 and Leu141 have been identified to be required for pKID to interact with KIX [40, 46]. A critical role of Leu141 has also been suggested based on the results of mutating KIX Tyr650 [38] and phi-value analysis obtained from Go-type modeling [17]. Moreover, our simulations confirm the important role of secondary structure formation in binding to KIX, as demonstrated by an NMR and biochemical study [38]. This study concluded that the minimal requirement for interaction (of KID, phosphorylated or un-phosphorylated, and C-Myb activation domains) with the native site of KIX is binding-coupled stabilization of an amphipathic helix (αB in pKID) [38]. Furthermore, our observation that a structural element of the αB helix forms initial native contacts during the binding is also consistent with other studies emphasizing the important role of the preformed structural elements in partner recognition by IDPs [47-49].

Based on our simulation data, we propose the following mechanism for the productive binding of pKID to KIX. In the highly disordered encounter complex (En1), a few native contacts begin to form between pKID αB residues (particularly by Leu138 and also by Ile137) and the KIX surface. Except for the formation of these binding contacts (present in up to 10% En1 complexes), Leu138 and Ile137 constitute the most helical region of pKID (helical in

about 40% En1 complexes). The key binding and folding step involves significant strengthening of the above-mentioned contacts, which is primarily enhanced by the formation of native contacts of Leu141, and also native contacts of Tyr134. As highlighted in the Results section (characteristics of the En2 complex), Leu141 seems to play a major anchoring role in binding to KIX, being also involved in αB helix formation. Overall, the key transition state involves a network of weak native interactions (present in 20-30% of En2) formed by the 134-141 fragment of pKID αB (helical in about 50% of En2). Following the formation of the transition state, the remaining native contacts may be formed.

Conclusion

In summary, the binding and folding scenario described above resembles the nucleation-condensation mechanism whose variations appear to describe the overall features of the folding of most protein domains [50-52]. A similar concept of the binding and folding scenario has been recently proposed based on protein engineering and kinetic experiments (phi value analysis) of the ACTR/NCBD system [45], or stopped-flow spectroscopy study of reaction between disordered peptides and PDZ domains [43]. Overall, it is likely that this mechanism is common for the coupled folding and binding of IDPs.

Acknowledgments

The authors acknowledge funding from Foundation for Polish Science TEAM project [TEAM/2011-7/6] co-financed by the EU European Regional Development Fund operated within the Innovative Economy Operational Program; Polish National Science Centre [NN301071140]; Polish Ministry of Science and Higher Education [IP2011 024371].

References

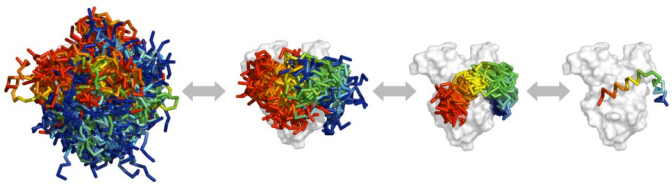
1. Koshland, D. E., *Application of a Theory of Enzyme Specificity to Protein Synthesis*. *Proc Natl Acad Sci U S A* 1958**,** *44*, 98-104.
2. Boehr, D. D.; Nussinov, R.; Wright, P. E., *The role of dynamic conformational ensembles in biomolecular recognition*. *Nature chemical biology* 2009**,** *5*, 789-96.
3. Hammes, G. G.; Chang, Y. C.; Oas, T. G., *Conformational selection or induced fit: a flux description of reaction mechanism*. *Proc Natl Acad Sci U S A* 2009**,** *106*, 13737-41.
4. Mittag, T.; Kay, L. E.; Forman-Kay, J. D., *Protein dynamics and conformational disorder in molecular recognition*. *Journal of Molecular Recognition* 2010**,** *23*, 105-116.
5. Uversky, V. N.; Gillespie, J. R.; Fink, A. L., *Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins: Structure, Function, and Bioinformatics* 2000**,** *41*, 415-427.

6. Wright, P. E.; Dyson, H. J., *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm*. *J Mol Biol* 1999**,** *293*, 321-31.
7. Eliezer, D., *Biophysical characterization of intrinsically disordered proteins*. *Current Opinion in Structural Biology* 2009**,** *19*, 23-30.
8. Rauscher, S.; Pomes, R., *Molecular simulations of protein disorder*. *Biochem Cell Biol* 2010**,** *88*, 269-90.
9. Uversky, V. N., *Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes*. *Chemical Society reviews* 2011**,** *40*, 1623-34.
10. Shoemaker, B. A.; Portman, J. J.; Wolynes, P. G., *Speeding molecular recognition by using the folding funnel: The fly-casting mechanism*. *Proceedings of the National Academy of Sciences* 2000**,** *97*, 8868-8873.
11. Vajda, S.; Kozakov, D., *Convergence and combination of methods in protein-protein docking*. *Curr Opin Struct Biol* 2009**,** *19*, 164-70.
12. Lensink, M. F.; Wodak, S. J., *Docking and scoring protein interactions: CAPRI 2009*. *Proteins: Structure, Function, and Bioinformatics* 2010**,** *78*, 3073-3084.
13. Sieradzan, A. K.; Liwo, A.; Hansmann, U. H., *Folding and self-assembly of a small protein complex*. *J Chem Theory Comput* 2012**,** *8*, 3416-3422.
14. Huang, Y.; Liu, Z., *Kinetic Advantage of Intrinsically Disordered Proteins in Coupled Folding–Binding Process: A Critical Assessment of the "Fly-Casting" Mechanism*. *Journal of molecular biology* 2009**,** *393*, 1143-1159.
15. Ganguly, D.; Chen, J., *Topology-based modeling of intrinsically disordered proteins: Balancing intrinsic folding and intermolecular interactions*. *Proteins: Structure, Function and Bioinformatics* 2011**,** *79*, 1251-1266.
16. Huang, Y.; Liu, Z., *Nonnative interactions in coupled folding and binding processes of intrinsically disordered proteins*. *PLoS One* 2010**,** *5*, e15375.
17. Turjanski, A. G.; Gutkind, J. S.; Best, R. B.; Hummer, G., *Binding-induced folding of a natively unstructured transcription factor*. *PLoS Comput Biol* 2008**,** *4*, e1000060.
18. Huang, Y.; Liu, Z., *Smoothing molecular interactions: the "kinetic buffer" effect of intrinsically disordered proteins*. *Proteins* 2010**,** *78*, 3251-9.
19. Dadarlat, V. M.; Skeel, R. D., *Dual role of protein phosphorylation in DNA activator/coactivator binding*. *Biophys J* 2011**,** *100*, 469-77.
20. Chen, H. F., *Molecular dynamics simulation of phosphorylated KID post-translational modification*. *PLoS ONE* 2009**,** *4*.
21. Espinoza-Fonseca, L. M., *Thermodynamic aspects of coupled binding and folding of an intrinsically disordered protein: a computational alanine scanning study*. *Biochemistry* 2009**,** *48*, 11332-4.
22. Kmiecik, S.; Gront, D.; Kolinski, A., *Towards the high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field*. *Bmc Struct Biol* 2007**,** *7*, 43.
23. Kolinski, A.; Bujnicki, J. M., *Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models*. *Proteins* 2005**,** *61 Suppl 7*, 84-90.
24. Blaszczyk, M.; Jamroz, M.; Kmiecik, S.; Kolinski, A., *CABS-fold: Server for the de novo and consensus-based prediction of protein structure*. *Nucleic Acids Res* 2013**,** *41*, W406-11.
25. Kmiecik, S.; Kolinski, A., *Characterization of protein-folding pathways by reduced-space modeling*. *Proc Natl Acad Sci U S A* 2007**,** *104*, 12330-5.
26. Kmiecik, S.; Kolinski, A., *Folding pathway of the B1 domain of protein G explored by multiscale modeling*. *Biophys J* 2008**,** *94*, 726-736.

27. Kmiecik, S.; Kolinski, A., *Simulation of chaperonin effect on protein folding: a shift from nucleation-condensation to framework mechanism*. *J Am Chem Soc* 2011**,** *133*, 10283-9.
28. Kmiecik, S.; Gront, D.; Kouza, M.; Kolinski, A., *From coarse-grained to atomic-level characterization of protein dynamics: transition state for the folding of B domain of protein A*. *J Phys Chem B* 2012**,** *116*, 7026-32.
29. Wabik, J.; Kmiecik, S.; Gront, D.; Kouza, M.; Koliński, A., *Combining Coarse-Grained Protein Models with Replica-Exchange All-Atom Molecular Dynamics*. *International Journal of Molecular Sciences* 2013**,** *14*, 9893-9905.
30. Jamroz, M.; Orozco, M.; Kolinski, A.; Kmiecik, S., *Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field*. *Journal of Chemical Theory and Computation* 2013**,** *9*, 119-125.
31. Jamroz, M.; Kolinski, A.; Kmiecik, S., *CABS-flex: Server for fast simulation of protein structure fluctuations*. *Nucleic Acids Res* 2013**,** *41*, W427-31.
32. Kurcinski, M.; Kolinski, A., *Steps towards flexible docking: modeling of three-dimensional structures of the nuclear receptors bound with peptide ligands mimicking co-activators' sequences*. *J Steroid Biochem Mol Biol* 2007**,** *103*, 357-60.
33. Kurcinski, M.; Kolinski, A., *Theoretical study of molecular mechanism of binding TRAP220 coactivator to Retinoid X Receptor alpha, activated by 9-cis retinoic acid*. *J Steroid Biochem Mol Biol* 2010**,** *121*, 124-9.
34. Steczkiewicz, K.; Zimmermann, M. T.; Kurcinski, M.; Lewis, B. A.; Dobbs, D.; Kloczkowski, A.; Jernigan, R. L.; Kolinski, A.; Ginalski, K., *Human telomerase model shows the role of the TEN domain in advancing the double helix for the next polymerization step*. *Proc Natl Acad Sci U S A* 2011**,** *108*, 9443-8.
35. Kolinski, A., *Protein modeling and structure prediction with a reduced representation*. *Acta biochimica Polonica* 2004**,** *51*, 349-71.
36. Gront, D.; Kmiecik, S.; Kolinski, A., *Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates*. *Journal of computational chemistry* 2007**,** *28*, 1593-7.
37. Sugase, K.; Dyson, H. J.; Wright, P. E., *Mechanism of coupled folding and binding of an intrinsically disordered protein*. *Nature* 2007**,** *447*, 1021-5.
38. Zor, T.; Mayr, B. M.; Dyson, H. J.; Montminy, M. R.; Wright, P. E., *Roles of phosphorylation and helix propensity in the binding of the KIX domain of CREB-binding protein by constitutive (c-Myb) and inducible (CREB) activators*. *The Journal of biological chemistry* 2002**,** *277*, 42241-8.
39. Radhakrishnan, I.; Perez-Alvarado, G. C.; Parker, D.; Dyson, H. J.; Montminy, M. R.; Wright, P. E., *Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions*. *Cell* 1997**,** *91*, 741-52.
40. Shaywitz, A. J.; Dove, S. L.; Kornhauser, J. M.; Hochschild, A.; Greenberg, M. E., *Magnitude of the CREB-dependent transcriptional response is determined by the strength of the interaction between the kinase-inducible domain of CREB and the KIX domain of CREB-binding protein*. *Molecular and cellular biology* 2000**,** *20*, 9409-22.
41. Gabdoulline, R. R.; Wade, R. C., *Biomolecular diffusional association*. *Curr Opin Struct Biol* 2002**,** *12*, 204-13.
42. Bachmann, A.; Wildemann, D.; Praetorius, F.; Fischer, G.; Kiefhaber, T., *Mapping backbone and side-chain interactions in the transition state of a coupled protein folding and binding reaction*. *Proc Natl Acad Sci U S A* 2011**,** *108*, 3952-7.
43. Haq, S. R.; Chi, C. N.; Bach, A.; Dogan, J.; Engstrom, A.; Hultqvist, G.; Karlsson, O. A.; Lundstrom, P.; Montemiglio, L. C.; Stromgaard, K.; Gianni, S.; Jemth, P., *Side-chain*

interactions form late and cooperatively in the binding reaction between disordered peptides and PDZ domains. *J Am Chem Soc* 2012**,** *134*, 599-605.
44. Karlsson, O. A.; Chi, C. N.; Engstrom, A.; Jemth, P., *The transition state of coupled folding and binding for a flexible beta-finger*. *J Mol Biol* 2012**,** *417*, 253-61.
45. Dogan, J.; Mu, X.; Engstrom, A.; Jemth, P., *The transition state structure for coupled binding and folding of disordered protein domains*. *Scientific reports* 2013**,** *3*, 2076.
46. Shih, H. M.; Goldman, P. S.; DeMaggio, A. J.; Hollenberg, S. M.; Goodman, R. H.; Hoekstra, M. F., *A positive genetic selection for disrupting protein-protein interactions: identification of CREB mutations that prevent association with the coactivator CBP*. *Proc Natl Acad Sci U S A* 1996**,** *93*, 13896-901.
47. Fuxreiter, M.; Simon, I.; Friedrich, P.; Tompa, P., *Preformed structural elements feature in partner recognition by intrinsically unstructured proteins*. *J Mol Biol* 2004**,** *338*, 1015-26.
48. Sivakolundu, S. G.; Bashford, D.; Kriwacki, R. W., *Disordered p27Kip1 exhibits intrinsic structure resembling the Cdk2/cyclin A-bound conformation*. *J Mol Biol* 2005**,** *353*, 1118-28.
49. Knott, M.; Best, R. B., *A preformed binding interface in the unbound ensemble of an intrinsically disordered protein: evidence from molecular simulations*. *PLoS Comput Biol* 2012**,** *8*, e1002605.
50. Fersht, A. R., *Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications*. *Proc Natl Acad Sci U S A* 1995**,** *92*, 10869-73.
51. Daggett, V.; Fersht, A. R., *Is there a unifying mechanism for protein folding? Trends in biochemical sciences* 2003**,** *28*, 18-25.
52. Itzhaki, L. S.; Otzen, D. E.; Fersht, A. R., *The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding*. *J Mol Biol* 1995**,** *254*, 260-88.

TOC graphic



Praca V (P.V)

Horwacik, Irena, **Mateusz Kurciński**, Małgorzata Bzowska, Aleksandra K Kowalczyk, Dominik Czaplicki, Andrzej Koliński, and Hanna Rokita. 2011. **Analysis and Optimization of Interactions Between Peptides Mimicking the GD2 Ganglioside and the Monoclonal Antibody 14G2a.** *International journal of molecular medicine* 28(1): 47–57.

Praca zawiera wyniki analizy eksperymentalnej i teoretycznej zjawiska strukturalnej mimikry molekularnej gangliozydu GD2 przez grupę precyzyjnie zaprojektowanych peptydów w kompleksie z przeciwciałem 14G2a. Nadekspresja gangliozydu GD2 skojarzona jest z występowaniem nerwiaka zarodkowego – groźnego nowotworu atakującego niemowlaki. Autorzy poszukują związków imitujących obecność w organizmie antygenu – potencjalnej szczepionki.

Wkład autora rozprawy polegał na przeprowadzeniu modelowania stabilności kompleksów przeciwciało–antygen dla różnych peptydów i wsparcie danych eksperymentalnych przy wyborze najlepszego kandydata do dalszych badań.

Analysis and optimization of interactions between peptides mimicking the GD2 ganglioside and the monoclonal antibody 14G2a

IRENA HORWACIK[1], MATEUSZ KURCIŃSKI[3], MAŁGORZATA BZOWSKA[2], ALEKSANDRA K. KOWALCZYK[1], DOMINIK CZAPLICKI[1], ANDRZEJ KOLIŃSKI[3] and HANNA ROKITA[1]

[1]Laboratory of Molecular Genetics and Virology, [2]Department of Immunology, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, 30-387 Krakow; [3]Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, 02-093 Warsaw, Poland

Received December 23, 2010; Accepted February 15, 2011

DOI: 10.3892/ijmm.2011.655

Abstract. Overexpression of the GD2 ganglioside (GD2) is a hallmark of neuroblastoma. The antigen is used in neuroblastoma diagnosis and to target newly developed therapies to cancer cells. Peptide mimetics are novel approaches in the design of antigens for vaccine development. We previously reported the isolation of five GD2-mimicking peptides from the LX-8 phage display library with the monoclonal antibody (mAb) 14G2a. The goal of our current study was to analyze and optimize the binding of the peptide mimetics to the mAb 14G2a. Therefore, we performed further experiments and supported them with molecular modeling to investigate structure-activity relationships that are the basis for the observed mimicry of GD2 by our peptides. Here, we show that the peptides have overlapping binding sites on the mAb, 14G2a and restricted specificity, as they did not crossreact with other ganglioside-specific antibodies tested. In addition we demonstrate that the phage environment was involved in the process of selection of our peptides. The AAEGD sequence taken from the viral major coat protein, p8, and added to the C-termini of the peptides #65, #85 and #94 significantly improved their binding to the mAb, 14G2a. By application of analogs with amino acid substitutions and sequence truncations, we elucidated the structure-activity relationships necessary for the interactions between the 14G2a mAb and the peptide #94 (RCNPNMEPPRCF). We identified amino acids indispensable for the observed GD2-mimicry by #94 and confirmed a pivotal role of the disulphide bridge between the cysteine residues of #94 for binding to the mAb 14G2a. More importantly, we report five new peptides demonstrating a significant improvement of mAb 14G2a binding. The experimental data were supported and expanded with molecular modeling tools. Taken together, the experimental results and the *in silico* data allowed us to probe in detail the mechanism of the molecular mimicry of GD2 by the peptides. Additionally, we significantly optimized binding of the leading peptide sequence #94 to the mAb 14G2a. We can conclude that our findings add to the knowledge on factors governing selections of peptide mimetics from phage-display libraries.

Correspondence to: Dr Irena Horwacik, Laboratory of Molecular Genetics and Virology, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, 7 Gronostajowa St., 30-387 Krakow, Poland
E-mail: irena.horwacik@uj.edu.pl

Key words: cancer, carbohydrate mimicking peptide, neuroblastoma, GD2 ganglioside

Introduction

Peptide libraries based on bacteriophage expression are pivotal tools to study the biomolecular interactions and are proved sources of novel ligands with potential therapeutic or biotechnological value (1). Prominent examples are surrogate antigens of both protein and non-protein molecules identified with antibodies. Such mimetic peptides can be applied in situations where the use of a native ligand may not be best suited (2,3). Isolation of mimetics is based on a phenomenon of molecular mimicry and is attributed to the conformational similarity between two or more otherwise chemically diverse molecules.

An important group of peptide mimetics consists of peptides developed for immunotherapy approaches to provide alternative antigens to replace glycans in vaccine formulations. Carbohydrates are relevant medical targets, because they are major constituents of the cell wall of bacterial or fungal pathogens, and antigens well-known for their aberrant expression in cancer. However, they are often poorly or non-immunogenic in infants, children, and immunocompromised individuals (4,5). Correlations between antigenic and immunogenic properties of carbohydrate-mimicking peptides have been investigated to show that peptide surrogates can successfully replace carbohydrate antigens (6-8). Additionally, the relative ease to manufacture and modify peptides should be highlighted, in contrast to the often challenging and costly synthesis of carbohydrates.

The GD2 ganglioside is an important example of a tumor-associated carbohydrate antigen. It is highly overexpressed in cancer cells of a childhood tumor, neuroblastoma, but also in other malignancies such as melanoma, sarcoma, or small cell lung carcinoma (9-11). GD2 is a glycolipid, consisting of an outer carbohydrate part, charged by the presence of two sialic acid residues, and a membrane anchor formed by a ceramide (fatty acid sphingosine).

Several GD2-binding mAbs have been developed (12,13). Such antibodies are currently used for the diagnosis and monitoring of treatment response in neuroblastoma patients. In clinical trials, anti-GD2 monoclonal antibody-based therapy of neuroblastoma patients was shown to be effective, and it is currently further refined to improve poor survival of patients with high risk disease (14,15). In addition to passive approaches, immunization strategies targeting GD2 are under development. One of the approaches tested is the active immunization with surrogates of GD2, which are exploited to improve the immunogenicity of GD2. They include anti-idiotypic antibodies (16,17) and peptide mimotopes (18,19). In practice, application of such surrogate antigens is often preceded by analysis of the molecular basis of mimicry, which then enables optimization of the leading sequence.

Recently, we have described a series of 12-amino acid constrained peptides capable of binding the anti-GD2 mAb, 14G2a, isolated from the LX-8 phage-displayed peptide library (20). The mAb, 14G2a, can effectively mediate *in vitro* and *in vivo* cytotoxicity against neuroblastoma cells (21) and has a well-established record of medical applications (22,23). The peptide sequences isolated by us with the mAb 14G2a can be grouped in two clusters: one consisting of the sequences #85, #D, #8, and the second including #65, #94 sequences (Table I). In competition tests against human neuroblastoma IMR-32 cells, which abundantly express the GD2 ganglioside, all peptides were shown to mimic the tumor-associated carbohydrate antigen.

The goal of our current study was to optimize the binding of the peptide mimetics to the mAb 14G2a. Therefore, we performed further *in vitro* experiments and supported them with molecular modeling to investigate the structure-activity relationships that are the basis for the observed mimicry of GD2 by the peptides. Here, as a result, we report five new peptide sequences with significantly improved binding to the mAb 14G2a. Additionally, we determined reactivity of the peptides with other ganglioside-specific antibodies and demonstrate that the peptides have overlapping binding sites on the mAb 14G2a. By design and analysis in the competition tests of several peptide analogs with amino acid substitutions, sequence truncations, and extensions we probed in detail the molecular mechanism of the GD2 mimicry observed for one of our peptides (#94).

Materials and methods

Cell lines, antibodies, gangliosides. IMR-32 (a GD2-positive human neuroblastoma cell line) (10) was purchased from ATCC (LCG Standards, Lomianki, Poland). The IMR-32 cells were grown in E-MEM supplemented with non-essential amino acids (1%), sodium pyruvate (1 mM), gentamicin (50 mg/l), 10% fetal bovine serum (FBS) (complete medium E-MEM). Murine mAb 14G2a (IgG2a), recognizing the GD2 ganglioside was produced by a hybridoma cell line that was a gift from Dr R.A. Reisfeld (The Scripps Institute, La Jolla, CA, USA). Hybridomas producing mouse mAbs recognizing GD2, i.e., mAb ME361-S2a (IgG2a) and mAb 126 (IgM) were purchased from ATCC. All hybridomas were grown in DMEM (with a glucose concentration of 4.5 g/l) supplemented with gentamicin (50 mg/l) and 10% FBS (complete DMEM). The cell lines were cultured at 37˚C in a humidified 5% CO_2 incubator. The cell culture reagents were purchased from Sigma-Aldrich (St. Louis, MO, USA). Other antibodies used in the study included: the GD3 ganglioside (GD3) binding mouse mAb ME3.6 (IgG3) (BD Biosciences Pharmingen, San Diego, CA, USA); the horseradish peroxidase (HRP)-conjugated, anti-p8 mouse mAb (Amersham Biosciences, Uppsala, Sweden); the fluorescein-5-isothiocyanate (FITC)-conjugated goat affinity purified $F(ab')_2$ fragments to mouse immunoglobulins IgA, IgG, IgM (Cappel, MP Biomedicals, Aurora, OH, USA); and the HRP-conjugated anti-mouse IgG (whole molecule) rabbit antibodies (Sigma-Aldrich). The gangliosides used in the study included: GD2 from human brain and GD3 from bovine butter milk (both form Calbiochem, Merck Chemicals Ltd., Nottingham, UK).

Peptides. Peptide sequences of GD2 mimetics (#8, #65, #85, #94, #D) expressed on phages and #0-phage (a phage without any fusion peptide) were isolated during biopanning of the LX-8 library with mAb 14G2a [for a detailed description of the biopanning and the preparation of the phage-expressed peptides see Horwacik *et al* (20)]. The major coat protein p8 is used for peptide display in the LX-8 library (24). Peptides were synthesized by GenScript USA Inc. (Piscataway, NJ, USA) and Sigma-Aldrich. The peptides were delivered in a lyophilized form at a purity of ~90%. To dissolve the lyophilized peptides, first N,N-dimethylformamide (Fluka) was added to each peptide sample, then diluted with water to a peptide concentration of 20 or 10 mM.

Production, purification, and modification of the mAb. The cell culture media (serum-free) containing monoclonal antibodies were sterile filtered through 0.45 μm syringe filters to remove any particulates. Two mouse IgG2a antibodies, the mAb, 14G2a, and the mAb, ME361-S2a, were purified with HiTrap protein G columns (Amersham Biosciences) according to the manufacturer's protocol. For mAb 126, a column with a protein L resin was used (Pierce, Thermo Scientific). The antibodies bound to the columns were eluted with 0.1 M glycine-HCl buffer (pH 2.7), and neutralized with 1 M Tris-HCl (pH 9.0). The fractions containing the antibodies were dialyzed against a final volume of 4 liters of phosphate buffered saline (PBS), pH 7.2, for 24 h at 4˚C. The concentrations of all the purified antibodies were determined with the Bicinchoninic assay (Sigma-Aldrich) using bovine serum albumin (BSA) as a standard (Sigma-Aldrich). The purity of the antbodies was verified using SDS-PAGE. Biotinylation of mAb 14G2a was performed with sulfo-NHS-LC biotin (Sigma-Aldrich) (25).

*Flow cytometry analyses of the free peptides binding to mAb, 14G2*a. The synthetic peptide mimetics of the GD2 ganglioside, their analogs with the designed sequence modifications,

and the control #T2 peptide with the HEDIISLWDQSL sequence derived from the HIV-1 envelope glycoprotein, (26) were incubated overnight at 4˚C with 200 ng of the 14G2a mAb in 2% BSA/PBS, (BSA from Amresco Inc., Solon, OH, USA). To determine the binding of the peptides to mAb 14G2a, the peptides were used over a range of concentrations obtained from serial dilutions decreasing by the factor of 2. The next day, the samples were used in a competition assay with GD2-positive IMR-32 cells for 1 h at 4˚C. After the washing steps, the binding of mAb 14G2a was detected with mouse Ig-specific FITC-conjugated goat $F(ab')_2$ fragments using flow cytometry (BD™ LSR II with BD FACSDiva software, BD Biosciences). Cells (10^4 per sample) were collected, and the signal from the cells stained with the secondary antibody alone was used to determine the positively stained pools of the IMR-32 cells. In these pools, the mean fluorescence intensity (MFI) and % of stained cells (%SC) were measured, and then used for the calculation of the fluorescence index (FLindex) according to the formula: FLindex = (MFI x %SC)/100. Values of the percentage of the inhibition were calculated using the formula: % inhibition = (1 - $FLindex_{peptide}/FLindex_{max}$) x 100%, where the $FLindex_{max}$ was calculated from samples with the the mAb 14G2a, alone. At least three experiments were performed for each of the analyzed sequences. Mean values of three independent experiments were presented on graphs with SEM (standard error of the mean) for the error bars. Statistical significance was verified with independent two sample t-tests with $p \leq 0.05$, $p \leq 0.01$, $p \leq 0.001$.

Competitive ELISA. Wells of the high protein binding plates were coated with 1 μg of streptavidin (Sigma-Aldrich) in TBS buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl) overnight at 4˚C. The next day, the wells were washed with TBS, and blocked with 2% BSA/TBS for 1 h at 37˚C. After 3 washes with 0.1% Tween-20/TBS, the biotinylated 14G2a mAb (0.1 μg/well) in 2% BSA/TBS was applied to the wells for 1 h at room temperature (RT). After 4 washes, the synthetic peptides #85 and #94 mimicking GD2 and the control non-binding peptide, #T2, all at a 0.3 mM concentration, were mixed with phages expressing peptides #65, #85, #94 (10^{11} phage particles per well) for 1.5 h at RT. Unbound phages were removed with 6 washes with 0.1% Tween-20/TBS, and then the mAb 14G2a-bound phages were detected with an anti-p8 HRP-conjugated mAb for 1 h at RT. After 6 washes with 0.1% Tween-20/TBS, the TMB substrate reagent set (BD Biosciences Pharmingen) was used to develop the signals. Absorbance values at 450 nm were collected. Samples were tested in duplicates. Three independent experiments were performed, and results from a representative experiment are presented as mean values (±SEM). Statistical significance was verified with independent two t-tests with $p \leq 0.05$, $p \leq 0.01$, $p \leq 0.001$.

*ELISA to detect binding of peptides to ganglioside-specific antibodies other than mAb 14G2*a. The peptides #65, #94, #85 expressed on the surface of phages were used in duplicates to coat 96-wells plates (MaxiSorp surface, NUNC A/S, Roskilde, Denmark) in TBS overnight at 4˚C, and then washed 3 times with TBS. Phage #0 (with no peptide expressed) was used as a negative control in the assay. The next day, the wells were blocked with 2% BSA/TBS for 1 h at 37˚C and then washed 4 times with 0.1% Tween-20/TBS. Incubations with anti-ganglioside antibodies (0.2 μg/well) were carried out for 1 h at RT, and then the wells were washed 4 times with 0.1% Tween-20/TBS. The binding of each antibody was detected for 1 h at RT with the secondary HRP-conjugated anti-mouse IgG antibodies diluted in 2% BSA/TBS. After the final washes of the wells, signals were developed with TMB Substrate, and then absorbance values at 450 nm were collected. To prove the specificity of the antibodies used in the assay, additional wells were coated with 50 ng of GD2 or GD3 in ethanol, and then left for the diluent to evaporate. Three independent experiments were performed and the results from a representative experiment are presented as mean values.

Molecular modeling. Interactions between mAb 14G2a and the investigated peptides were modeled in the following procedure. First, template-based modeling was done to obtain a three-dimensional structure of the receptor (14G2a). Blast scanning of the PDB (Protein Data Bank) database against the receptor sequence identified the 1SVZ structure as a template for comparative modeling. Blast-generated alignment was used. Structure was modeled with CABS, a reduced space molecular modeling tool, which has been described in great detail elsewhere (27,28). Next, the obtained model of the receptor was used in docking simulations of five investigated peptides. Docking was carried out by CABS Dock, a CABS-based tool enhanced with multi-chain protein support. Each of the simulations produced 5,000 models of the receptor-peptide complex, reduced to a C-α trace representation. These models were subsequently clustered with respect to their mutual similarity measured as root mean square deviation (RMSD) computed on C-α atoms. For each of the peptides, the structure closest to the centroid of the most populated cluster was selected for further processing, which included reconstruction of the C-α only to full-atom models and energy minimization by means of molecular mechanics. These were performed with the SYBYL Amber99 force field with default parameters. Models of the complexes were ranked by the interaction energy between the receptor and five investigated peptides. The same all-atom models were used to identify the interacting amino acid residues.

Results

*The mimicking peptides #65, #85 and #94 share overlapping binding areas on the anti-GD2 mAb, 14G2*a. We aimed at choosing one leading peptide sequence for further optimization of mAb 14G2a binding. Therefore, we investigated structure-activity relationships that are necessary for the observed mimicry of GD2. The peptides #8 and #D were not included in the assays, due to their high sequence similarity with the peptide #85 (Table I). Also, in respect to peptide #D a significantly lower binding to the antibody was observed, which prompted us to exclude the peptide sequence from further analyses. As for peptides #65, #85 and #94 isolated with mAb 14G2a, we examined whether the mimetics share binding areas on the antibody rather than interacting with independent binding sites. We set up a competition ELISA,

Table I. Amino acid sequences of peptides mimicking the GD2 ganglioside isolated with mAb 14G2a from the LX-8 phage-display library.

Peptide code	Peptide sequence
#8	NCDLLTGPMLCV
#85	VCNPLTGALLCS
#D	GCDALSGHLLCS
#65	SCQSTRMDPNCW
#94	RCNPNMEPPRCF

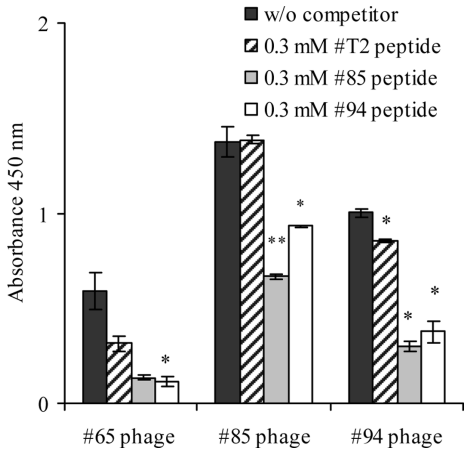


Figure 1. Results of the competitive ELISA. The assay was used to detect binding of #65, #85 and #94 peptides expressed on phages to the biotinylated mAb 14G2a, attached to streptavidin-coated wells in the presence of competitors (0.3 mM), i.e., free peptides #85, #94 and #T2. Binding was detected with the HRP-conjugated anti-p8 antibody and signals were developed with the TMB substrate, followed by measurement of the absorbance values at 450 nm. Data from a representative experiment of three separate experiments are shown (samples were tested in duplicates, mean values ± SEM are shown in the graph); $^{*}p\leq0.05$, $^{**}p\leq0.01$ (two sample independent t-tests).

in which the free peptides #85 and #94, and the control mAb 14G2a non-binding peptide #T2 (all used at a 0.3 mM concentration), were mixed with phages expressing peptides #65, #85 or #94, and were then allowed to compete for binding to the biotinylated-mAb 14G2a molecules attached to streptavidin-coated wells. Next, the viral particles interacting with the immobilized antibody were detected with an anti-p8 protein HRP-conjugated antibody.

Data from a representative inhibition experiment are shown in Fig. 1. Significant reductions of mAb 14G2a binding for the phages expressing the peptides #65, #85 and #94 were observed in the presence of the free GD2-mimetics #85 or #94 used as competitors, as compared to the control peptide #T2. The synthetic peptide #85 reduced the binding of the #65 phage to 43%, #85 phage to 48% and #94-phage to 35% of the signal measured in the presence of the #T2 peptide. For the synthetic peptide #94 the calculated values were 36% for the #65 phage, 67% for the #85 phage and 44% for the #94 phage. The effect was tested for two additional concentrations of the synthetic competitors (0.1 and 0.5 mM), and was observed to be dose-dependent (data not shown). The above data allowed us to conclude that despite clear differences in their amino acid sequences, the peptides #65, #94 and #85 may have overlapping binding sites on the 14G2a antibody molecule.

Peptides #65, #85 and #94 do not crossreact with other ganglioside-specific antibodies tested. Next, we decided to determine whether our peptides can bind to other commercially available GD2 ganglioside-specific monoclonal antibodies by ELISA. In addition to mAb 14G2a (used as a positive control), we tested two other mouse mAb binding to GD2, i.e., mAb ME361-S2a and mAb 126. A mouse anti-GD3 mAb (mAb ME3.6) was used as a specificity control. The phages expressing peptides #65, #85, #94 and the control phage #0 were used to coat wells of high protein binding plates, as the synthetic peptides were unable to efficiently adhere to the surface (data not shown). Additionally, the binding of all tested antibodies was positively verified in wells coated with GD2 or GD3. Results of the ELISA are presented in Table II. The phage expressing the peptides #65, #85 and #94 showed a diverse degree of binding to mAb 14G2a (a feature previously reported). However, we detected no binding of the peptides to any other antibodies than to mAb 14G2a. Also, phage #0 did not significantly bind to any of the antibodies tested. In a separate set of experiments, we observed that the peptides did not compete with GD2-positive IMR-32 cells for binding to mAb 126 and mAb ME361-S2a (data not shown). The results prompted us to conclude that the isolated peptides are mimicking a unique GD2 ganglioside epitope, as they specifically interact only with the paratope of the mAb 14G2a.

*Extension of GD2-mimicking peptides with amino acids of the phage coat protein, p8, significantly improves their binding to mAb 14G2*a. Aiming to optimize the binding of our

Table II. Cross reactivity of phage-expressed peptides #65, #85, #94 and #0 phage (a negative control) with ganglioside-specific antibodies measured by ELISA.

Antibody	$A_{450\ nm}$					
	#65	#85	#94	#0	GD2	GD3
Anti-GD2 mAb 14G2a	1.2	2.57	1.49	0.07	3.06	0.05
Anti-GD2 mAb 126	0.07	0.07	0.07	0.07	2.84	0.06
Anti-GD2 mAb ME361-S2a	0.09	0.09	0.09	0.09	1.576	0.09
Anti-GD3 mAb ME3.6	0.06	0.06	0.05	0.07	0.04	0.38

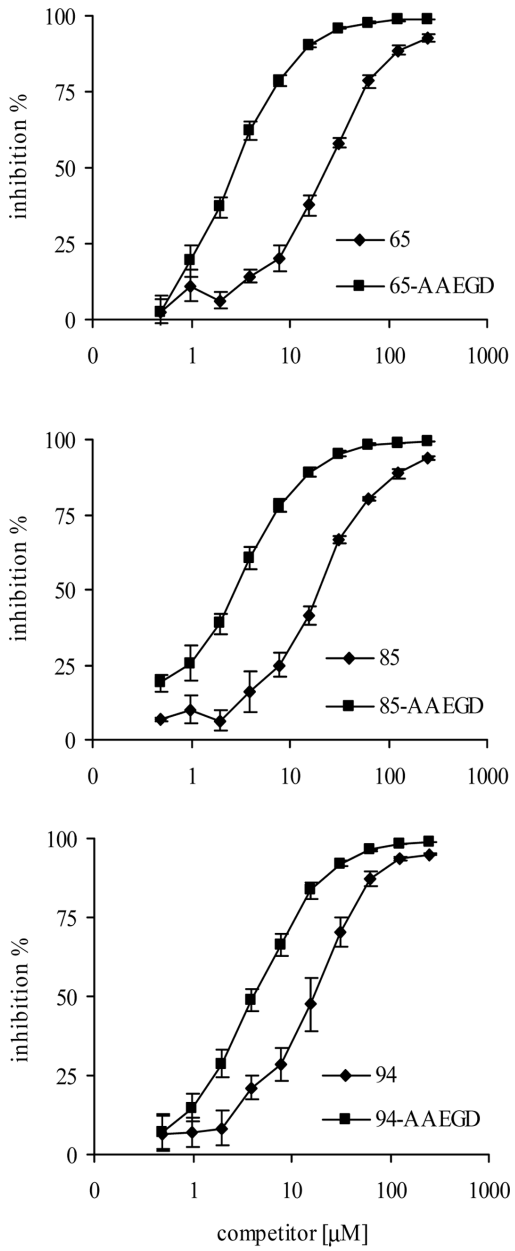


Figure 2. The results from competition experiments with #65, #85 and #94 and their modifications #65-AAEGD, #85-AAEGD, #94-AAEGD. Changes in mAb 14G2a binding after introduction of replacements were tested in the competition assay against the IMR-32 cells for the peptide concentration range of 0.49-250 μM using flow cytometry. Mean values ± SEM from four independent experiment are shown.

peptides to mAb 14G2a, we performed experiments to verify if the phage environment can influence their interaction with mAb 14G2a. Our hypothesis was supported by the observation that the first four or five N-terminal amino acids of the p8 are flexible, in the otherwise rigid structure formed by the interacting p8 monomers of the viral capsid (29). Therefore, we synthesized extended peptides by addition of the AAEGD sequence from the p8 protein used for display to the C-termini of the peptides #65, #85 and #94.

The results of the performed competition experiments against the GD2 positive IMR-32 neuroblastoma cells for both the initial and the extended peptides are presented in Fig. 2. For all the modified peptides a significant increase in binding to mAb 14G2a was measured as compared to the initially identified sequences. IC_{50} values (half maximal inhibitory concentration) for all 6 peptides tested are summarized in Table III. The IC_{50} value of the #65-AAEGD decreased 8-fold ($p \leq 0.01$) as compared to #65. Similarly, the IC_{50} value of the #85-AAEGD decreased 7-fold ($p \leq 0.001$) as compared to #85. Finally, a nearly 4-fold drop in the IC_{50} value was calculated for the #94-AAEGD in comparison to the initial #94 peptide ($p \leq 0.05$). The results allowed us to conclude that during the process of isolation of our peptides from the LX-8 phage-display library the presence of the viral p8 protein to some extent influenced the interaction between the peptides and the mAb 14G2a.

Alanine scanning of the GD2-mimicking peptide #94 allows identification of the structural basis necessary for the mAb 14G2a binding. Based on the experiments described above, we chose peptide #94 as the leading sequence for further analyses and optimization. In the next step, we performed site-specific mutagenesis of the initial 12-amino acid sequences of the #94 peptide by replacing its consecutive side chains with a small alanine. We evaluated the obtained analogs in competition assays for binding to the mAb 14G2a against GD2 in IMR-32 cells. The results from a representative experiment are shown in Fig. 3. For all except one of the alanine-substituted peptides, a significant decrease in the 14G2a-binding properties was observed in the assay. The substitutions with alanine of C-2, N-3, M-6, E-7, C-11 in the chain of the #94 peptide resulted in complete loss of binding to mAb 14G2a. The IC_{50} values could not be calculated for the analogs. This indicates their particularly critical role in

Table III. Results of extension experiments with the AAEGD sequence.

Peptide code	Peptide sequence[a]	IC_{50} (μM)[b]
#65	SCQSTRMDPNCW	25.0±2.2
#65-AAEGD	SCQSTRMDPNCWAAEGD	3.0±0.2
#85	VCNPLTGALLCS	20.4±1.6
#85-AAEGD	VCNPLTGALLCSAAEGD	3.1±0.3
#94	RCNPNMEPPRCF	17.1±2.8
#94-AAEGD	RCNPNMEPPRCFAAEGD	4.7±0.5

[a]Position of each sequence modification is underlined. [b]Half maximal inhibition concentration (mean values ± SEM from four independent experiments).

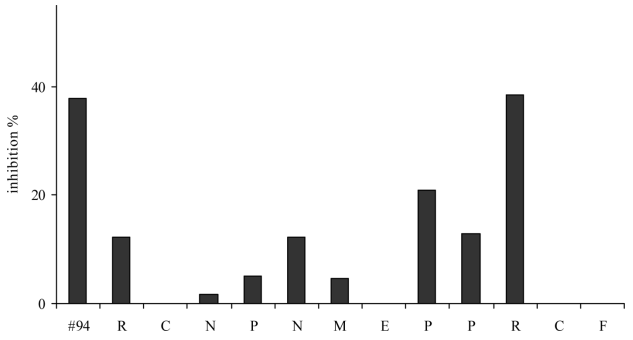


Figure 3. Identification of residues of the peptide #94 critical for the mAb 14G2a binding by alanine scanning. Changes in mAb 14G2a binding after introduction of alanine in place of the subsequent amino acid were tested in the competition assay against the IMR-32 cells using flow cytometry. Data for the 15.6 μM concentration of peptides are shown (for a representative of three separate experiments).

in interaction with mAb 14G2a. For the substitution of the P-4 and F-12 residues with alanine the values of IC_{50} were calculated, respectively, as 150.8 and 151.2 μM, while in the experiment the IC_{50} for #94 was 20.7 μM. Moreover, when R-1, N-5, P-8, P-9 were replaced with alanine the IC_{50} values were calculated from 45.4 to 56.8 μM. Only for the peptide in which R-10 was replaced by alanine we did not observe significant changes in the 14G2a-binding properties.

*Substitutions introduced into the #94 peptide allowed identification of a new sequence with higher binding to mAb 14G2*a. We further extended our replacement analysis based on the findings from the alanine scanning. We designed and analyzed additional analogs of the #94 sequence to explore the importance of key side-chain groups, i.e., N-3, M-6, E-7, F-12 and their respective length, hydrophobic or hydrophilic properties on the binding to mAb 14G2a. We substituted N-3 with Q or D, M-6 with L or F, E-7 with D, Q or N, and finally F-12 with 2 other natural aromatic amino acids, i.e., W or Y. Results of the competition assay for binding to mAb 14G2a, between such mutated peptides and GD2 expressed in IMR-32 cells, are shown in Fig. 4, and in Table IV. For both N-3 and E-7 all tested replacements in the initial #94 peptide chain resulted in a complete loss of the binding to mAb 14G2a

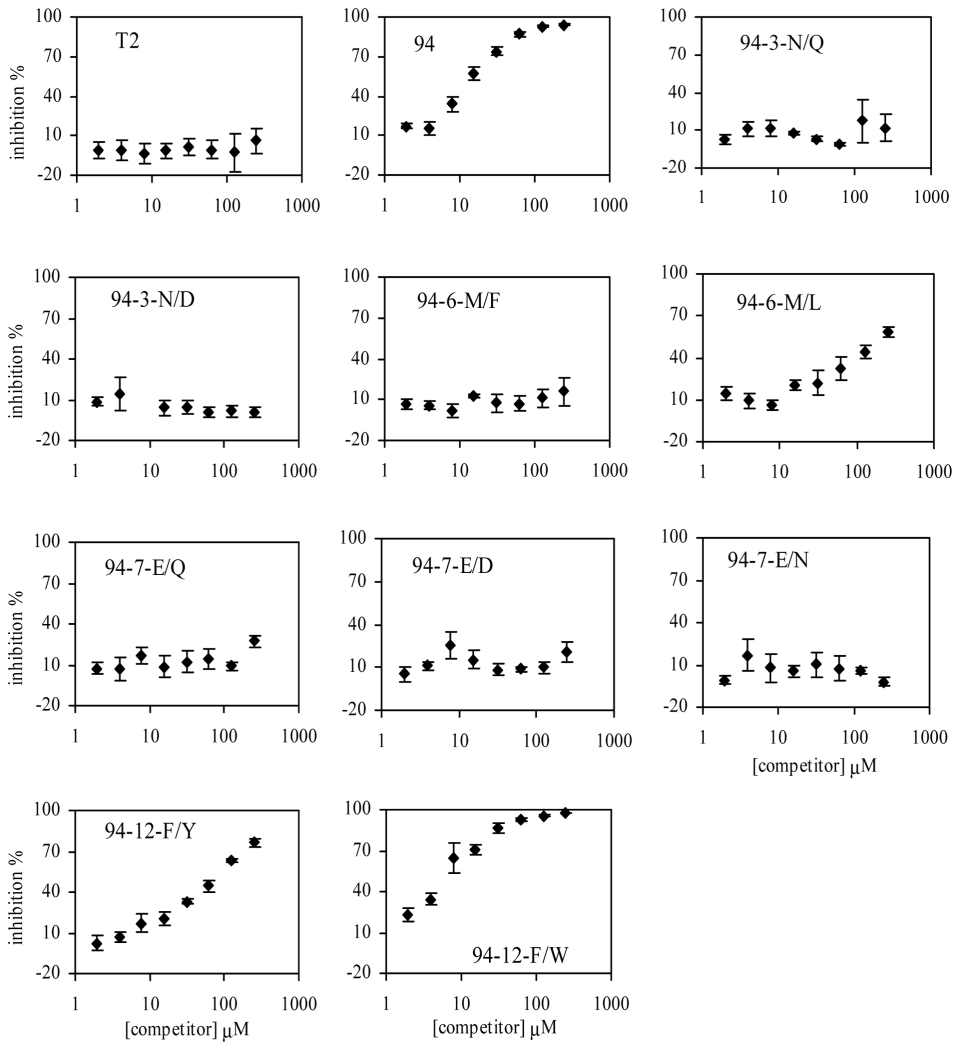


Figure 4. The results from competition experiments with #94, #T2 (a negative control) and #94 analogs containing substitutions of N-3 (N with Q or D), M-6, (M with L or F) E-7 (E with Q, D or N) and F-12 (F with Y or W). Changes in mAb 14G2a binding after introduction of replacements were tested in the competition assay against the IMR-32 cells for peptide concentrations from 1.95-250 μM using flow cytometry (the graphs with #94 and #T2 peptides were included for comparison). Mean values ± SEM from three independent experiments are shown.

Table IV. Results of substitution, truncation and elongation experiments within the #94 sequence.

Peptide code	Peptide sequence[a]	IC_{50} (μM)[b]
#T2	HEDIISLWDQSL	ni
#94	RCNPNMEPPRCF	14.0±2.2
#94-3-N/Q	RCQPNMEPPRCF	ni
#94-3-N/D	RCDPNMEPPRCF	ni
#94-6-M/F	RCNPNFEPPRCF	ni
#94-6-M/L	RCNPNLEPPRCF	136.4±33.5
#94-7-E/Q	RCNPNMQPPRCF	ni
#94-7-E/D	RCNPNMDPPRCF	ni
#94-7-E/N	RCNPNMNPPRCF	ni
#94-12-F/Y	RCNPNMEPPRCY	65.0±3.7
#94-12-F/W	RCNPNMEPPRCW	6.8±0.4
#94-Δ-1	--NPNMEP----	ni
#94-Δ-2	--NPNMEPPR--	ni
#94-Δ-3	-CNPNMEPPRCF	ni
#94-Δ-4	RCNPNMEPP-CF	115.9±16.8
#94-12-F/W-AAEGD	RCNPNMEPPRCWAAEGD	0.8±0.1

[a]The position of each mutation is underlined. [b]Mean values ± SEM from three independent experiments are shown. ni, no inhibition measured within the tested range of dilutions of the peptides.

and no IC_{50} values were observed. Likewise, the #94-6-M/F surrogate did not bind to the antibody. On the other hand, the replacement of M-6 with L did not completely abrogate binding of the peptide to mAb 14G2a, although a significant 10-fold rise in the mean IC_{50} value was observed for the peptide (IC_{50} = 136.4±33.5 μM, p≤0.05).

The analysis of the #94-12-F/Y sequence also showed a significant decrease of binding affinity to mAb 14G2a of the peptide with the introduced hydroxyl group to the aromatic chain with Y (a nearly 5-fold increase in the IC_{50} value was observed for the peptide, IC_{50} = 65.0±3.7 μM, p≤0.001). On the contrary, for #94-12-F/W, the peptide with a larger aromatic group of W introduced in the place of F, the new IC_{50} value was calculated to be 6.8±0.4 μM, 2-fold lower than for peptide #94 (p≤0.05).

Truncations of peptide #94 can significantly change binding to mAb 14G2a and confirm the pivotal role of the disulphide bridge in the interaction. To find smaller fragments of the #94 sequence that would bind to mAb 14G2a four shorter peptides derived from #94 were analyzed. They included two linear peptide sequences NPNMEP (6-amino acid long, #94-Δ1) and NPNMEPPR (8-amino acid long, #94-Δ2), as well as two 11-amino acid long peptides missing, respectively, R-1 (#94-Δ3) and R-10 (#94-Δ4). The results of the competition experiments are shown in Fig. 5. Both peptides lacking the flanking residues including the two C residues at positions 2 and 11 (NPNMEP and NPNMEPPR) did not bind to the receptor at all, despite the fact that the sequences still contained some of the amino acids pivotal for binding mAb 14G2a (Table IV). The results emphasize the importance of the disulphide bridge-forming cysteines in the observed binding process and confirm the alanine scanning data. Also, deletion of the R residue at position 1 (#94-Δ3) resulted in loss of binding to mAb 14G2a, throughout the entire range of concentrations tested. However, the peptide with the eliminated R at position 10 (#94-Δ4) could still bind to mAb 14G2a, even though a significant reduction of binding was observed (an 8-fold increase in IC_{50} was measured, IC_{50}=115.9±16.8 μM, p≤0.05, Table IV).

Peptide analog #94 combining the 12-F/W substitution with the C-terminal AAEGD-extension shows further improvement of mAb 14G2a binding. We demonstrated that the addition of the five amino acids of the p8 protein at the C-terminus of #94 improved mAb 14G2a-binding (Table III). Therefore, we evaluated one additional peptide (#94-12-F/W-AAEGD) containing both the 12-F/W substitution and the C-terminal AAEGD-extension and compared it to the initial #94 and the #94-AAEGD sequences. We observed a significant improvement of mAb 14G2a binding of #94-12-F/W-AAEGD, as compared to #94 and #94-AAEGD in the competition experiment (Fig. 6). To calculate the IC_{50} value for the #94-12-F/W-AAEGD, we measured the inhibition of #94-12-F/W-AAEGD within the 1.95-250 μM range of concentrations and compared it to #94. The value of IC_{50} was significantly improved for #94-12-F/W-AAEGD peptides as compared to #94, and was nearly 18-fold lower than #94 (IC_{50}=0.8±0.1 μM, p≤0.01, Table IV).

Molecular modeling data are in parallel with experimental results. Along with the *in vitro* experiments, we applied an *in silico* approach to gather information about structures of the peptide mimetics of GD2 bound to mAb 14G2a (Fig. 7). We calculated the free binding energies for the peptides binding to mAb 14G2a. The lowest binding energy was calculated

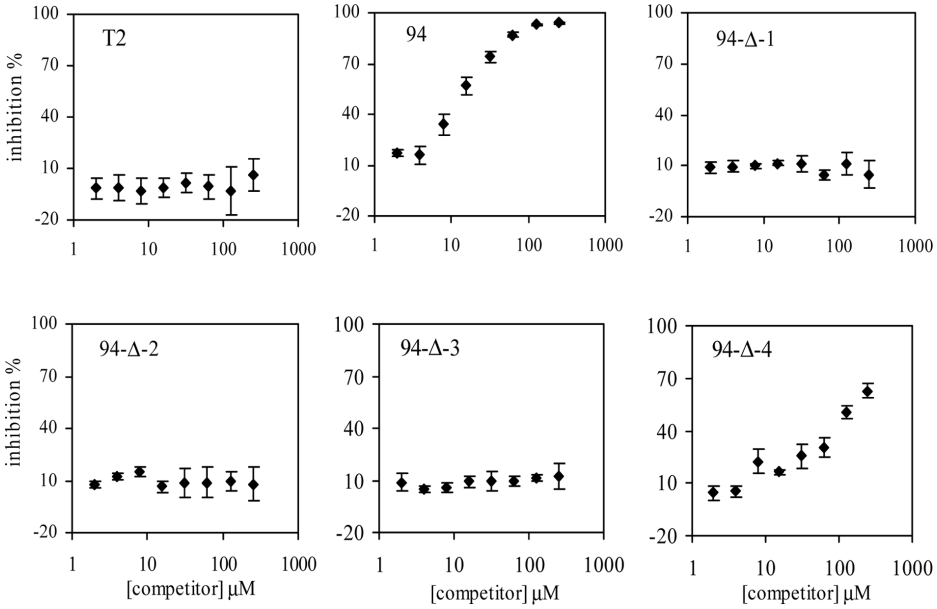


Figure 5. Results from experiments with #94, #T2 (a negative control) and #94 analogs bearing truncations of the peptide sequence. Changes in mAb 14G2a binding after introduction of modifications were tested in the competition assay against IMR-32 cells for peptide concentrations from 1.95-250 μM using flow cytometry (the graphs with #94 and #T2 peptides were included for comparison). Mean values ± SEM from three independent experiments are shown.

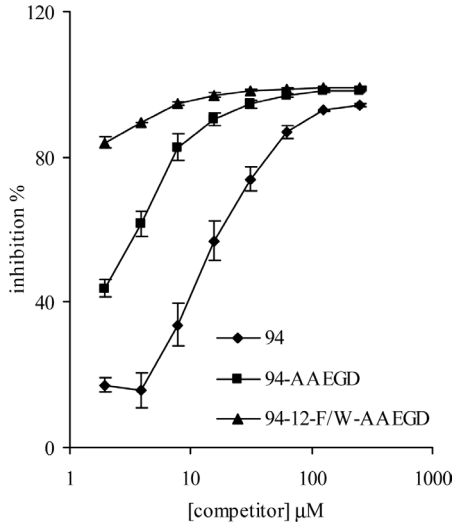


Figure 6. Results from experiments with peptides #94, #94-AAEGD and #94-12-F/W-AAEGD. Changes in mAb 14G2a binding after introduction of the modifications were examined in the competition assay against IMR-32 cells with peptide concentrations from 1.95-250 μM using flow cytometry. Mean values ± SEM from three independent experiments are shown.

for peptide #65 (-228.9 kcal/mol), while the highest was determined for peptide #D (-126.7 kcal/mol). For the rest of the peptides the calculated values were comparable, -170.5, -165.1 and -178.9 kcal/mol for #8, #85 and #94, respectively.

Analysis of the molecular model allowed us to identify amino acid residues of peptide #94 and the mAb 14G2a fragment engaged in the binding. Several close interactions (around 3Å distance) were identified between the amino acids of peptide #94 ligands and the mAb 14G2a (Table V). Analysis of the models built for peptides #8, #65, #85, #94 and #D and their interactions with the mAb 14G2a fragment support our experimental findings that the peptides share binding areas in the antibody paratope (data not shown). Our studies demonstrated that the computational approach and the experimental methods may successfully complement one another.

Discussion

Our research stems from the observation that binding of an antigen to its cognate antibody is rarely exclusive, and cross-reactivity of immunoglobulin molecules with other epitopes can often be detected (30). We applied this information to isolate peptides mimicking GD2 ganglioside that is one of the antigens used to target neuroblastoma cells. The mimicry phenomenon for peptides isolated from phage-display libraries can be investigated on many levels including phage-expressed and free peptides, or peptides fused to proteins (31,32). Finally, characterization of functional (agonist or antagonist activity) or immunological mimicry of peptides can be performed (33,34). We aimed to probe the molecular basis of mAb 14G2a binding by peptides isolated from the LX-8 library. As a result, we significantly improved binding to mAb 14G2a for the leading sequence (#94) with application of peptide engineering.

Here, we described several new findings unfolding the molecular basis for the observed interactions between the mAb 14G2a receptor and the peptide ligands mimicking GD2. Peptides #65, #85 and #94 were shown to occupy overlapping binding areas in the paratope of mAb 14G2a. A similar feature was reported for two peptides binding to the gp120 protein of HIV-1 that were isolated from phage-display peptide libraries (35). Our result can also be supported with molecular modeling data showing that peptides #8, #65, #85, #94 and #D are in close contact with overlapping amino acids from the modeled mAb 14G2a fragment. In separate experiments, we did not detect any binding of peptides #65, #85 and #94 to two additional GD2-specific antibodies, mAb ME361-S2a and mAb 126. This may suggest that the isolated peptides

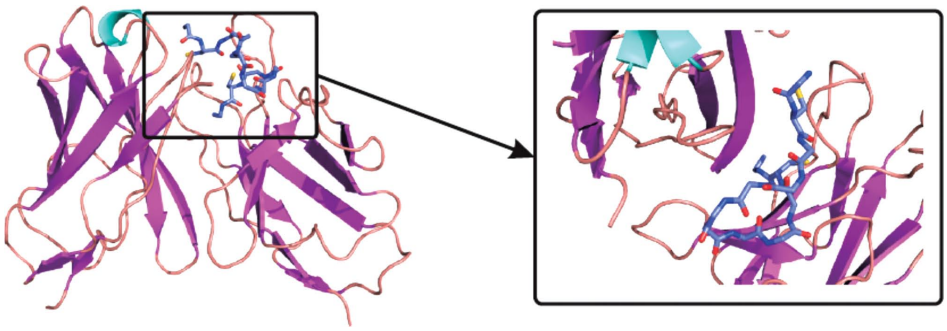


Figure 7. A model of #94 interaction with a fragment of mAb 14G2a.

Table V. Residues of the receptor (mAb 14G2a) and the ligand molecule (peptide #94) interacting in the molecular modeling experiments.

Receptor	Ligand	Distance (Å)
V_L: H-39	R-10	3.313
V_L: L-51	R-10	3.440
V_L: H-54	P-8	3.155
V_L: H-54	R-10	3.071
V_L: P-60	R-10	3.418
V_H: N-35	R-1	3.031
V_H: N-35	C-2	3.339
V_H: V-37	R-1	3.250
V_H: V-37	F-12	3.381
V_H: W-47	R-1	2.799
V_H: D-52	N-3	2.712
V_H: S-59	N-3	2.836
V_H: Y-94	F-12	3.465
V_H: V-96	F-12	3.254
V_H: E-100	E-7	3.157
V_H: E-100	P-9	3.495
V_H: Y-101	C-2	3.190
V_H: Y-101	E-7	3.153
V_H: Y-101	P-9	3.113
V_H: W-102	C-11	3.040

V_L, light chain fragment of the receptor; V_H, heavy chain of the receptor.

mimic a unique GD2 epitope. Another explanation could be that their reactivity patterns are restricted as they are specific only to mAb 14G2a that was used for their isolation from the LX-8 phage-display library. This can further be supported by observations of Harris *et al* who studied peptides isolated with a panel of mAb binding to a similar epitope of cell wall polysaccharide of group A *Staphylococcus*. They reported that the peptides bound only to the mAb used for their isolation (36). We can thus conclude that it might be beneficial to isolate peptide mimetics with other GD2-specific antibodies than mAb 14G2a. Combining such peptides mimicking different GD2 epitopes into polyepitope vaccines could broaden the spectrum of induced immune responses against GD2. Peptide mimetics binding to mAb ME361 have been reported (37). Another interesting approach to obtain peptide sequences is the *de novo* design of mimetics. Tong *et al* reported recently that peptidic GD2 ligands were rationally designed based on NMR and molecular modeling analyses of GD2 in free form and bound to the anti-GD2 mAb 3F8 (38).

Predicting whether high or low affinity peptides will be isolated from a library seems to be impossible. For peptides mimicking discontinuous epitopes or non-protein antigens, not only the linear sequence, but also the conformation of both the peptide and the binding site on the cognate antibody may play a role, to a various extent, in the isolation process, and later in the reactivity of the peptides (39,40). Additionally, the theoretical diversity of peptide libraries is reduced during their preparation steps and the number of peptide copies expressed on the surface of the viral particles may vary, which in turn affects the results of the screening process (41). Therefore, as a rule the initial leading sequences isolated from phage-display libraries are not the most optimal and require further optimization.

Here, we report five peptide sequences with significantly improved binding to mAb 14G2a. Three sequences arose from the C-terminal extension of #65, #85 and #94 with the AAEGD sequence from the p8 protein used for peptide-display. This finding adds to the knowledge on factors driving peptide selections from phage-display libraries. Such positive influence of the viral environment on the binding properties of phage-expressed peptides has also been reported by others (33,42). Therefore, we postulate that addition of amino acids adjacent to expressed peptides can be viewed as a possible mean to improve their interactions with the receptors used in screening of the phage-display libraries.

Two additional improved mimetics were obtained from the application of a systemic approach to analyze in detail the binding of one member of the identified GD2-mimicking peptide set, #94. We report that both sequence requirements and structure constrains are responsible for the antigenic reactivity with mAb 14G2a. Also, we identified residues critical for the binding. Based on the findings, we substituted F-12 with a larger W moiety and improved binding of the obtained analog of #94 to mAb 14G2a. Furthermore, introduction of such a larger aromatic group combined with the AAEGD-extension yielded the best of our peptides. Noteworthy, the tryptophan residue also occupies the last position in the #65 sequence, stressing that both shared and unique determinants of binding

of the peptides to mAb 14G2a can be identified. Altogether the data highlight the importance of the C-terminal part of #94 for the binding.

Finally, we were able to confirm the particularly important role of the disulphide bridge, formed by the C residue at positions 2 and 11, in the observed binding of the free peptides. The finding was previously reported for all our phage-display peptides based on the observed abrogation of their mAb 14G2a binding in the presence of two agents, dithiothreitol and N-ethylmaleimide, reducing and preventing them from reforming a disulphide bond (20). Altogether, for both the free and the phage-displayed peptides, we can assume that cyclic peptides could adopt or maintain the antigenically reactive structure necessary for GD2 mimicry. In the literature such observations have been reported, e.g., for the EMP1 mimetic of the hormone, erythropoietin, where the presence of a disulphide bridge at positions 6 and 15 was essential in both the 20-amino acid long original EMP1 sequence, as well as in the 13-amino acid long minimal active sequence. The bridge could not be substituted with an amide bond formed between the introduced side chains of glutamic acid at position 6 and lysine at the position 15 (43). Furthermore, introduction of sequence cyclization to an initially linear peptide was reported to positively modulate antigen reactivity (44).

In conclusion, the described results allowed us to gain a detailed insight into the mode of interaction between our GD2 mimicking peptides and mAb 14G2a. Additionally, we supported the experiments with an *in silico* approach that yielded data in agreement with *in vitro* results. Therefore, the methods used to construct the computer model were correct and possibly could be used in further studies on GD2 mimicry. Our findings widen the knowledge about factors governing selections of peptides from phage-display libraries. In the future, we plan to extend our research on the peptides mimicking GD2, as five new and stronger 14G2a-binding peptides were identified. We plan to further analyze the mimetics and correlate the observed antigenicity of the peptides with their possible immunogenicity, as they can be tested as antigens to induce GD2-specific immune responses. We are encouraged to continue research on peptides mimicking GD2, which is a clinically important tumor-associated carbohydrate antigen and a therapeutically relevant target for anticancer treatments.

Acknowledgements

We are grateful to Dr R.A. Reisfeld for the 14G2a hybridoma cell line. The study was mainly financed in years 2006-2010 from the research grant no. N302 034 31/3063 from the Polish Ministry of Science and Higher Education (to I.H.) and also partially from the Jagiellonian University Grants no. WBBB 8 (to H.R.) and no. BW37/137 (to I.H.).

References

1. Smith GP and Petrenko VA: Phage display. Chem Rev 97: 391-410, 1997.
2. Beenhouwer DO, May RJ, Valadon P and Scharff MD: High affinity mimotope of the polysaccharide capsule of *Cryptococcus neoformans* identified from an evolutionary phage peptide library. J Immunol 169: 6992-6999, 2002.
3. Torregrossa P, Buhl L, Bancila M, Durbec P, Schafer C, Schachner M and Rougon G: Selection of poly-alpha 2,8-sialic acid mimotopes from a random phage peptide library and analysis of their bioactivity. J Biol Chem 279: 30707-30714, 2004.
4. Fleuridor R, Lyles RH and Pirofski L: Quantitative and qualitative differences in the serum antibody profiles of human immunodeficiency virus-infected persons with and without *Cryptococcus neoformans* meningitis. J Infect Dis 180: 1526-1535, 1999.
5. Lucas AH, Rittenhouse-Olson K, Kronenberg M, Apicella MA, Wang D, Schreiber JR and Taylor CE: Carbohydrate moieties as vaccine candidates: meeting summary. Vaccine 28: 1121-1131, 2010.
6. Valadon P, Nussbaum G, Oh J and Scharff MD: Aspects of antigen mimicry revealed by immunization with a peptide mimetic of Cryptococcus neoformans polysaccharide. J Immunol 161: 1829-1836, 1998.
7. Cunto-Amesty G, Luo P, Monzavi-Karbassi B, Lees A and Kieber-Emmons T: Exploiting molecular mimicry to broaden the immune response to carbohydrate antigens for vaccine development. Vaccine 19: 2361-2368, 2001.
8. Shin JS, Lin JS, Anderson PW, Insel RA and Nahm MH: Monoclonal antibodies specific for Neisseria meningitidis group B polysaccharide and their peptide mimotopes. Infect Immun 69: 3335-3342, 2001.
9. Yoshida S, Fukumoto S, Kawaguchi H, Sato S, Ueda R and Furukawa K: Ganglioside G(D2) in small cell lung cancer cell lines: enhancement of cell proliferation and mediation of apoptosis. Cancer Res 61: 4244-4252, 2001.
10. Hettmer S, Ladisch S and Kaucic K: Low complex ganglioside expression characterizes human neuroblastoma cell lines. Cancer Lett 225: 141-149, 2005.
11. Ravindranath MH, Muthugounder S and Presser N: Ganglioside signatures of primary and nodal metastatic melanoma cell lines from the same patient. Melanoma Res 18: 47-55, 2008.
12. Cheung NK, Lazarus H, Miraldi FD, Abramowsky CR, Kallick S, Saarinen UM, Spitzer T, Strandjord SE, Coccia PF and Berger NA: Ganglioside GD2 specific monoclonal antibody 3F8: a phase I study in patients with neuroblastoma and malignant melanoma. J Clin Oncol 5: 1430-1440, 1987.
13. Kawashima I, Tada N, Fujimori T and Tai T: Monoclonal antibodies to disialogangliosides: characterization of antibody-mediated cytotoxicity against human melanoma and neuroblastoma cells *in vitro*. J Biochem 108: 109-115, 1990.
14. Choi BS, Sondel PM, Hank JA, Schalch H, Gan J, King DM, Kendra K, Mahvi D, Lee LY, Kim K and Albertini MR: Phase I trial of combined treatment with ch14.18 and R24 monoclonal antibodies and interleukin-2 for patients with melanoma or sarcoma. Cancer Immunol Immunother 55: 761-774, 2006.
15. Modak S and Cheung NK: Neuroblastoma: therapeutic strategies for a clinical enigma. Cancer Treat Rev 36: 307-317, 2010.
16. Zeytin HE, Tripathi PK, Bhattacharya-Chatterjee M, Foon KA and Chatterjee SK: Construction and characterization of DNA vaccines encoding the single-chain variable fragment of the anti-idiotype antibody 1A7 mimicking the tumor-associated antigen disialoganglioside GD2. Cancer Gene Ther 7: 1426-1436, 2000.
17. Basak S, Birebent B, Purev E, Somasundaram R, Maruyama H, Zaloudik J, Swoboda R, Strittmatter W, Li W, Luckenbach A, Song H, Li J, Sproesser K, Guerry D, Nair S, Furukawa K and Herlyn D: Induction of cellular immunity by anti-idiotypic antibodies mimicking GD2 ganglioside. Cancer Immunol Immunother 52: 145-154, 2003.
18. Bolesta E, Kowalczyk A, Wierzbicki A, Rotkiewicz P, Bambach B, Tsao CY, Horwacik I, Kolinski A, Rokita H, Brecher M, Wang X, Ferrone S and Kozbor D: DNA vaccine expressing the mimotope of GD2 ganglioside induces protective GD2 cross-reactive antibody responses. Cancer Res 65: 3410-3418, 2005.
19. Fest S, Huebener N, Weixler S, Bleeke M, Zeng Y, Strandsby A, Volkmer-Engert R, Landgraf C, Gaedicke G, Riemer AB, Michalsky E, Jaeger IS, Preissner R, Förster-Wald E, Jensen-Jarolim E and Lode HN: Characterization of GD2 peptide mimotope DNA vaccines effective against spontaneous neuroblastoma metastases. Cancer Res 66: 10567-10575, 2006.
20. Horwacik I, Czaplicki D, Talarek K, Kowalczyk A, Bolesta E, Kozbor D and Rokita H: Selection of novel peptide mimics of the GD2 ganglioside from a constrained phage-displayed peptide library. Int J Mol Med 19: 829-839, 2007.
21. Mujoo K, Kipps TJ, Yang HM, Cheresh DA, Wargalla U, Sander DJ and Reisfeld RA: Functional properties and effect on growth suppression of human neuroblastoma tumors by isotype switch variants of monoclonal antiganglioside GD2 antibody 14.18. Cancer Res 49: 2857-2861, 1989.

22. Lode HN, Handgretinger R, Schuermann U, Seitz G, Klingebiel T, Niethammer D and Beck J: Detection of neuroblastoma cells in $CD34^+$ selected peripheral stem cells using a combination of tyrosine hydroxylase nested RT-PCR and anti-ganglioside GD2 immunocytochemistry. Eur J Cancer 33: 2024-2030, 1997.
23. Cheung NK, Kushner BH and Kramer K: Monoclonal antibody-based therapy of neuroblastoma. Hematol Oncol Clin North Am 15: 853-866, 2001.
24. Bonnycastle LL, Mehroke JS, Rashed M, Gong X and Scott JK: Probing the basis of antibody reactivity with a panel of constrained peptide libraries displayed by filamentous phage. J Mol Biol 258: 747-762, 1996.
25. Menendez A, Bonnycastle LLC, Pan OCC and Scott JK: Screening peptide libraries. In: Phage Display. A Laboratory Manual. 1st edition. Barbas CF, Burton DR, Scott JK and Silverman GJ (eds.) Cold Spring Harbor Laboratory Press, New York, pp17.8-17.11, 2001.
26. Kröpelin M, Süsal C, Daniel V and Opelz G: Inhibition of HIV-1 rgp120 binding to $CD4^+$ T cells by monoclonal antibodies directed against the gp120 C1 or C4 region. Immunol Lett 63: 19-25, 1998.
27. Koliński A: Protein modeling and structure prediction with a reduced representation. Acta Biochim Pol 51: 349-371, 2004.
28. Kurciński M and Koliński A: Steps towards flexible docking: modeling of three-dimensional structures of the nuclear receptors bound with peptide ligands mimicking co-activators' sequences. J Steroid Biochem Mol Biol 103: 357-360, 2007.
29. Colnago LA, Valentine KG and Opella SJ: Dynamics of fd coat protein in the bacteriophage. Biochemistry 26: 847-854, 1987.
30. van Regenmortel MH: From absolute to exquisite specificity. Reflections on the fuzzy nature of species, specificity and antigenic sites. J Immunol Methods 216: 37-48, 1998.
31. Zwick MB, Bonnycastle LL, Noren KA, Venturini S, Leong E, Barbas CF III, Noren CJ and Scott JK: The maltose-binding protein as a scaffold for monovalent display of peptides derived from phage libraries. Anal Biochem 264: 87-97, 1998.
32. Popkov M, Sidrac-Ghali S, Alakhov V and Mandeville R: Epitope-specific antibody response to HT-1080 fibrosarcoma cells by mimotope immunization. Clin Cancer Res 6: 3629-3635, 2000.
33. Venkatesh N, Im SH, Balass M, Fuchs S and Katchalski-Katzir E: Prevention of passively transferred experimental autoimmune myasthenia gravis by a phage library-derived cyclic peptide. Proc Natl Acad Sci USA 97: 761-766, 2000.
34. Pameijer CR, Navanjo A, Meechoovet B, Wagner JR, Aguilar B, Wright CL, Chang WC, Brown CE and Jensen MC: Conversion of a tumor-binding peptide identified by phage display to a functional chimeric T cell antigen receptor. Cancer Gene Ther 14: 91-97, 2007.
35. Ferrer M and Harrison SC: Peptide ligands to human immunodeficiency virus type 1 gp120 identified from phage display libraries. J Virol 73: 5795-5802, 1999.
36. Harris SL, Craig L, Mehroke JS, Rashed M, Zwick MB, Kenar K, Toone EJ, Greenspan N, Auzanneau FI, Marino-Albernas JR, Pinto BM and Scott JK: Exploring the basis of peptide-carbohydrate crossreactivity: evidence for discrimination by peptides between closely related anti-carbohydrate antibodies. Proc Natl Acad Sci USA 94: 2454-2459, 1997.
37. Wondimu A, Zhang T, Kieber-Emmons T, Gimotty P, Sproesser K, Somasundaram R, Ferrone S, Tsao CY and Herlyn D: Peptides mimicking GD2 ganglioside elicit cellular, humoral and tumor-protective immune responses in mice. Cancer Immunol Immunother 7: 1079-1089, 2008.
38. Tong W, Gagnon M, Sprules T, Gilbert M, Chowdhury S, Meerovitch K, Hansford K, Purisima EO, Blankenship JW, Cheung NK, Gehring K, Lubell WD and Saragovi HU: Small-molecule ligands of GD2 ganglioside, designed from NMR studies, exhibit induced-fit binding and bioactivity. Chem Biol 17: 183-194, 2010.
39. Ferrières G, Villard S, Pugnière M, Mani JC, Navarro-Teulon I, Rharbaoui F, Laune D, Loret E, Pau B and Granier C: Affinity for the cognate monoclonal antibody of synthetic peptides derived from selection by phage display. Role of sequences flanking the binding motif. Eur J Biochem 267: 1819-1829, 2000.
40. Jung HH, Yi HJ, Lee SK, Lee JY, Jung HJ, Yang ST, Eu YJ, Im SH and Kim JI: Structural analysis of immunotherapeutic peptides for autoimmune Myasthenia gravis. Biochemistry 46: 14987-14995, 2007.
41. Malik P, Terry TD, Gowda LR, Langara A, Petukhov SA, Symmons MF, Welsh LC, Marvin DA and Perham RN: Role of capsid structure and membrane protein processing in determining the size and copy number of peptides displayed on the major coat protein of filamentous bacteriophage. J Mol Biol 260: 9-21, 1996.
42. Dorgham K, Dogan I, Bitton N, Parizot C, Cardona V, Debré P, Hartley O and Gorochov G: Immunogenicity of HIV type 1 gp120 CD4 binding site phage mimotopes. AIDS Res Hum Retroviruses 21: 82-92, 2005.
43. Johnson DL, Farrell FX, Barbone FP, McMahon FJ, Tullai J, Hoey K, Livnah O, Wrighton NC, Middleton SA, Loughney DA, Stura EA, Dower WJ, Mulcahy LS, Wilson IA and Jolliffe LK: Identification of a 13-amino acid peptide mimetic of erythropoietin and description of amino acids critical for the mimetic activity of EMP1. Biochemistry 37: 3699-3710, 1998.
44. Gomes P, Giralt E and Andreu D: Antigenicity modulation upon peptide cyclization: application to the GH loop of foot-and-mouth disease virus strain C1-Barcelona. Vaccine 19: 3459-3466, 2001.

Praca VI (P.VI)

Steczkiewicz, Kamil, Michael T Zimmermann, **Mateusz Kurciński**, Benjamin A Lewis, Drena Dobbs, Andrzej Kloczkowski, Robert L Jernigan, Andrzej Koliński and Krzysztof Ginalski. 2011. **Human Telomerase Model Shows the Role of the TEN Domain in Advancing the Double Helix for the Next Polymerization Step.** *Proceedings of the National Academy of Sciences of the United States of America* 108(23): 9443–48.

Autorzy przeprowadzili wieloskalowe modelowanie struktury teoretycznego modelu ludzkiej telomerazy. W pracy zastosowano i połączono różnorodne techniki takie jak modelowanie porównawcze, dokowanie molekularne, analizę drgań normalnych.

Wkład autora rozprawy obejmował przeprowadzenie serii symulacji dokowania za pomocą programu FTDOCK, a następnie sporządzenie rankingu otrzymanych modeli na postawie ich oceny w polu siłowym programu CABS.

Human telomerase model shows the role of the TEN domain in advancing the double helix for the next polymerization step

Kamil Steczkiewicz[a], Michael T. Zimmermann[b,c], Mateusz Kurcinski[d], Benjamin A. Lewis[c,e], Drena Dobbs[c,e], Andrzej Kloczkowski[b,f], Robert L. Jernigan[b,c,1], Andrzej Kolinski[d], and Krzysztof Ginalski[a]

[a]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland; [b]Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011; [c]Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011; [d]Faculty of Chemistry, University of Warsaw, Warsaw, Poland; [e]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011; and [f]Battelle Center for Mathematical Medicine, Research Institute at Nationwide Children's Hospital and Department of Pediatrics, Ohio State University College of Medicine, Columbus, OH 43205

Edited by Robert Baldwin, Stanford University, Stanford, CA, and approved April 22, 2011 (received for review October 13, 2010)

Telomerases constitute a group of specialized ribonucleoprotein enzymes that remediate chromosomal shrinkage resulting from the "end-replication" problem. Defects in telomere length regulation are associated with several diseases as well as with aging and cancer. Despite significant progress in understanding the roles of telomerase, the complete structure of the human telomerase enzyme bound to telomeric DNA remains elusive, with the detailed molecular mechanism of telomere elongation still unknown. By application of computational methods for distant homology detection, comparative modeling, and molecular docking, guided by available experimental data, we have generated a three-dimensional structural model of a partial telomerase elongation complex composed of three essential protein domains bound to a single-stranded telomeric DNA sequence in the form of a heteroduplex with the template region of the human RNA subunit, TER. This model provides a structural mechanism for the processivity of telomerase and offers new insights into elongation. We conclude that the RNA:DNA heteroduplex is constrained by the telomerase TEN domain through repeated extension cycles and that the TEN domain controls the process by moving the template ahead one base at a time by translation and rotation of the double helix. The RNA region directly following the template can bind complementarily to the newly synthesized telomeric DNA, while the template itself is reused in the telomerase active site during the next reaction cycle. This first structural model of the human telomerase enzyme provides many details of the molecular mechanism of telomerase and immediately provides an important target for rational drug design.

polymerase | protein motions | structure prediction

Telomerases are essential for maintaining chromosome length and integrity (1–3). They complement the cellular DNA-dependent DNA polymerase replication machinery that is not capable of fully replicating chromosomal ends, leading to telomeric DNA sequence erosion. Loss of telomerase activity is tolerated to some extent in yeast, worm, plant, and mouse (4), but after a few generations, telomeres typically become too short to perform their essential functions. Excessive telomere shortening can result in chromosome degradation, illegitimate recombination and end-to-end fusion, the compromise of cell cycle regulation, and, ultimately, cell death (5, 6). Most eukaryotic organisms utilize telomerases for the successive synthesis and maintenance of telomeric DNA repeats at chromosome ends, replenishing the capability for further cell proliferation in stem cell lineages (7). Interestingly, the activity of telomerases in fully differentiated somatic cells is strongly down-regulated over time, in concert with aging, and a direct correlation between telomere shortening and aging has been demonstrated (8). In human cells, increased telomerase activity can increase renewal capacity in certain tissues, which has been interpreted as delaying the aging process. However, pathogenic overexpression of telomerases is also a hallmark of many human cancers (9), and several studies have shown that telomerase malfunction can lead to diseases in humans (4, 10), including dyskeratosis congenita (11). Thus, telomerase appears to be a key player critical in maintaining the balance between normal cellular differentiation (and aging) and the aberrant proliferation manifested in carcinogenic transformation (and immortality).

Influences of changes in telomerase activity have been observed in many biological processes not directly related to telomere maintenance (12, 13). For example, Gonzalez-Suarez et al. found that induced somatic expression of telomerase led to increased cellular proliferation and growth and, consequently, enhanced wound healing in mice (14). Studies on promyelocytic leukemia cells revealed that telomerase expression may also inhibit apoptosis (15). When overexpressed, telomerase is directed to the mitochondria and appears to help protect cells from H_2O_2-mediated damage (16). Recently, Blackburn and colleagues have shown that changes in telomerase activity are associated with human stress-related syndromes, including major depression (17, 18).

Telomerases function as specialized reverse transcriptases (19), RNA-dependent DNA polymerases capable of synthesizing multiple copies of the telomeric DNA repeat sequence by using an intrinsic RNA template to direct telomeric DNA synthesis (1, 13). Telomere repeats are often shown as 5′-$(TTAGGG)_n$-3′, which is the DNA repeat unit. In this work we choose to focus on the RNA template, displaying its coding region as 5′-UAACCC-3′. For clarity, the alignment region 3′-AUC-5′ of hTR pairs to the DNA primer sequence 5′-TAG-3′ in order to synthesize the next DNA addition. The newly synthesized telomeric DNA repeats are added to the overhanging single-stranded 3′ end of the DNA at the chromosome termini. In other respects, the telomerase reverse transcriptase mechanism appears to be similar to that of well-studied retroviral reverse transcriptases (1, 12, 20).

The human telomerase enzyme contains a template-encoding RNA molecule, TER (TElomerase RNA or hTR) and a primary protein component, TERT (TElomerase Reverse Transcriptase) with several functional domains: TEN (Telomerase Essential

Author contributions: D.D., A. Kloczkowski, R.L.J., A. Kolinski, and K.G. designed research; K.S., M.T.Z., M.K., and B.A.L. performed research; K.S., M.T.Z., M.K., D.D., A. Kloczkowski, R.L.J., A. Kolinski, and K.G. analyzed data; and K.S., M.T.Z., D.D., A. Kloczkowski, R.L.J., A. Kolinski, and K.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed: E-mail: jernigan@iastate.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1015399108/-/DCSupplemental.

N-terminal domain), TRBD (Telomerase RNA Binding Domain), RT (Reverse Transcriptase domain), and the C-Terminal Extension (CTE) (1, 12, 13). The RT domain can be further divided into two distinguishable subdomains: the "fingers" involved in nucleotide binding and processivity and the "palm" providing the polymerase catalytic residues and DNA primer grip. The C-terminal extension is responsible for interaction with DNA and has been proposed to correspond to the RT "thumb" domain (20). Because this CTE appears to be structurally equivalent to the C-terminal α-helical thumb of retroviral RT (21), it will be referred to as the RT thumb domain below. In describing the human TERT structure here, we thus refer to RT as fingers, palm, and thumb for clarity and simplicity. Both RT and TER are essential for telomerase activity, together forming the active site that catalyzes deoxynucleotide addition. The TRBD domain links the RT and TER components, and the TEN domain is proposed to facilitate the repetitive repeat addition mode of telomerases, which is one of the distinguishing features of telomerases, relative to classical reverse transcriptases (1, 12). Efficient repeat addition processivity is governed by multiple mechanisms involving both TER (22) and protein subdomains of TERT (23), as well as additional telomerase-associated processivity factors, notably TPP1-POT1 in humans, which enhances telomerase processivity by slowing primer dislocation and facilitating translocation (24). The TEN domain contains an "anchor" site (25) that is thought to help stabilize the bound single-stranded telomeric DNA substrate within the complex, while the intrinsic RNA template is realigned for the next, iterative reverse transcription cycle. Therefore, the complex is capable of processively synthesizing a long array of single-stranded telomeric DNA repeats by repeatedly copying the 6-nt long RNA template region within the TER component (1, 12). In addition to processivity factors, several species-specific accessory proteins are critical for telomerase assembly, subcellular localization, and function in vivo (12, 26). In human cell lines, for example, the catalytically active form of telomerase includes dyskerin (27), which together with NHP2 and NOP10, is required for stability and accumulation of the RNA component of human telomerase in vivo (28).

Recently, X-ray structures of the full length *T. castaneum* telomerase (containing RT and TRBD domains) alone (PDB ID codes 3DU5 and 3DU6) (21) and in complex with an RNA:DNA hairpin (PDB ID code 3KYL) (29) have been published. In addition, the crystal structures for separate TRBD (PDB ID code 2R4G) (30) and TEN (PDB ID code 2B2A) (31) domains from *T. thermophila* are now available. Despite much experimental effort, the detailed molecular mechanism of human telomerase enzymatic activity and the structural details of the interactions between the TEN domain and the other components of the complex (RT, TRBD, TER, and the telomeric DNA) are still not fully known. The absence of a high-resolution experimentally determined structure for the assembled telomerase core catalytic complex is a serious impediment to designing experiments that could further elucidate the molecular mechanism of telomerase action. Moreover, species-specific features of telomerase structure and function make obtaining a complete structure of the human telomerase enzyme particularly important. Thus we employ here a theoretical modeling approach to generate the entire 3D structure of the human TERT, TEN, and TRBD bound to a DNA substrate and its RNA template.

Automatic homology modeling using available web-based servers was not feasible because the amino acid sequence identities between the human telomerase domains and the corresponding RT and TRBD structures in the Protein Data Bank (PDB) are only 24% and 22%, respectively. This level of sequence similarity is near the limit for the reliable use of standard homology detection methods based on PSI-BLAST or RPS-BLAST. Furthermore, in our hands, such standard sequence comparison methods were unable to detect significant sequence similarity between the N-terminal domain of the human telomerase protein and the TEN domain of *Tetrahymena*, or any other known protein structure. To obviate this problem, we used an advanced meta-profile comparison method, Meta-BASIC (32), to map the human telomerase protein sequence onto sequences of the determined structures from *Tribolium* and *Tetrahymena*. The mappings obtained were confirmed by using a variety of fold recognition methods. Together with detailed manual inspection, these approaches allow us to generate highly accurate sequence-to-structure alignments between the human telomerase sequence and relevant structural templates. We then built three-dimensional models separately for TEN and the other components of the human telomerase complex including a hybrid RNA:DNA double helix formed between the RNA template and the single-stranded telomeric DNA substrate and assembled them by using protein–protein docking, guided by relevant experimental data. Based on the resulting structural model of the telomerase enzyme, we propose a mechanism for human telomerase action in which interactions between TEN and the RT:TRBD subcomplex play a critical role, and where the elongating telomeric DNA is stabilized by the TEN domain. We hypothesize that the helical structure of the heteroduplex formed between the RNA template and the telomeric DNA substrate is actively maintained during the individual repeated telomerase reactions producing a single copy of the template. Following this, the RNA template must be repositioned relative to the active site. We propose that its translocation proceeds along the extending helix due to constraints imposed by the TEN "anchoring" domain.

Results and Discussion

The present work was motivated by the lack of a complete structural model that could explain the detailed functional molecular mechanism of human telomerase. By using distant homology detection, comparative structural modeling, and computational docking, we developed a model of the human telomerase complex (Fig. 1). Then by using elastic network models, we investigated the intrinsic motions of the modeled structure. Our goals were twofold: (*i*) to understand how the individual telomerase protein domains and the intrinsic TER component interact in the assembled human telomerase RNP enzyme, and (*ii*) to generate a model illustrating how the telomerase RNP enzyme binds to and extends single-stranded telomeric DNA by reverse transcription of telomeric repeat sequences.

Several studies indicate that the TEN domain functions as an "anchor" to bind and stabilize the telomeric DNA substrate, contributing to the processivity of the repetitive reverse transcriptase activity (13, 23, 25). However, no structure-based explanation for how TEN contributes to the overall function of the telomerase enzyme or its processivity has been proposed previously. Here we suggest that TEN plays a critical role in controlling the processive step in which telomerase advances by one base on the strand being copied. Similarly, many experiments have determined the effects of specific amino acid substitutions or deletions on telomerase enzymatic activity and have provided us with useful information about key residues (see Table S1). It is important to reconcile these with a structural model and a structural mechanism. To date, however, it has not been possible to understand the effects of these changes due to the lack of a complete structural model for the telomerase RNP. The present structure will facilitate such investigations.

To derive a 3D model for human telomerase enzyme we used the available structures for telomerase components from *T. castaneum* and *T. thermophila*, including the recently released structure for RT and TRBD domains solved with RNA template and telomeric DNA (29). The modeled RNA:DNA heteroduplex was extended as observed in the closely related HIV-1 reverse transcriptase structure (33) to construct interactions with the TEN domain (see *Methods*). The human TEN domain, which

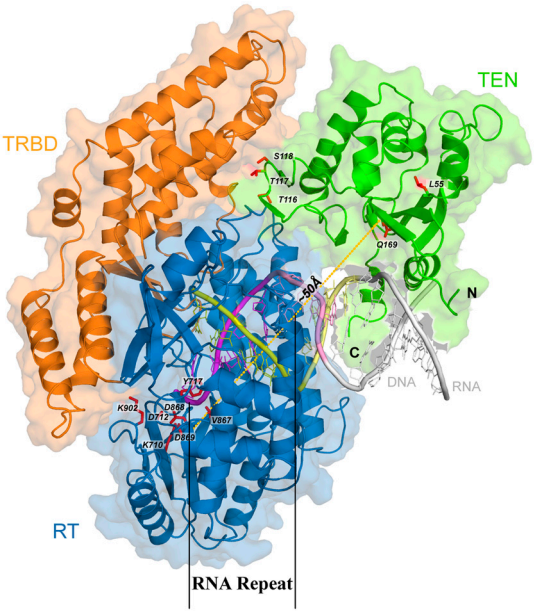


Fig. 1. Partial model of human telomerase ribonucleoprotein complex. The model includes the TERT protein component composed of the catalytic reverse transcriptase domain (RT), the RNA binding domain (TRBD), and the N-terminal "anchor" domain (TEN), bound to a heteroduplex formed by oligomers corresponding to the template-encoding RNA (TER) and the single-stranded human telomeric DNA substrate. The template region (5′-UAACCC-3′) of the RNA is shown in yellow (labeled as RNA repeat) and the complementary region of DNA in magenta, while the partial template repeat (5′-UAAC-3′) in the RNA is shown in lighter yellow, and its complement in the DNA in pink. The N- and C-terminal α-helices of TEN interact with the major groove of the heteroduplex. Residues with experimentally determined influence on telomerase function or assembly are labeled and shown in red. The orange dotted line shows the approximately 50- Å distance between TEN (Q169) and the RT active site (D869).

was modeled separately, seems to be designed to accommodate a nucleic acid double helix, as does the corresponding TEN structure from *T. thermophila* (31). It has a well-defined cleft to interact with the phosphate backbone and two helical segments that fit well into the major groove of the double helix. We used molecular docking to assemble the entire complex by fitting the TEN model into a position that ensured appropriate interaction with both the telomeric DNA and the other telomerase protein domains. We also considered available data suggesting plausible RNA binding sites, conservation of surface residues, and certain amino acid mutations that have been shown to impact the enzymatic activity (see references in Table S1). The final model shows that the central ring-shaped part of the human telomerase structure, formed by TRBD and RT domains, accommodates the RNA:DNA heteroduplex and provides catalytic residues (Fig. S1). The ring is coupled with the TEN domain, which interacts with TRBD, the RT thumb, and the RNA:DNA major groove and helps stabilize the substrate within the active site during repeated reaction cycles.

Reverse Transcriptase and Nucleic Acid Binding Domains. The RT and TRBD domains form the central, ring-shaped core of the telomerase RNP complex (Fig. S1). They provide the catalytic activity of the enzyme by bringing together the necessary active site residues. Previous analyses of sequence conservation within these domains highlighted several motifs, shared by the majority of reverse transcriptases, including telomerases (1, 34). As shown in Fig. S2 starting from the N terminus, these motifs are: CP, T, 1, 2, 3, A, B′, C, D, and E. Motifs CP and T belong to the α-helical TRBD domain (Fig. S1), which mediates interactions between the template-carrying RNA component and the telomerase. The CP and T motifs have been shown to be directly involved in RNA binding (12, 35) and are required for the proper assembly of the complete telomerase ribonucleoprotein complex (13, 36). Motifs A and C in the RT palm contain the conserved active site signature sequence, $KXD(X)_nDD$, in which three invariant aspartic acid residues (D712, D868, and D869) coordinate two Mg^{2+} ions, while the lysine (K710) provides the base for the deoxynucleotide condensation reaction (13, 21, 34) (see Fig. S3*A*).

The pocket surrounding the catalytic amino acids is lined with several residues that help position deoxynucleotide substrates (A, T, and G) with respect to the complementary RNA template-encoded ribonucleotides in the active site, as in the *T. castaneum* structure (21, 29). Three conserved uncharged residues—Y717, Q833, and V867 (from motifs A, B′, and C, respectively)—form a hydrophobic pocket adjacent to the catalytic aspartates and take part in nucleotide binding (29). Residue V867 has been shown to alter human telomerase substrate specificity (37), and residue Q833 corresponds to Q151 in HIV-1 reverse transcriptase, where mutations cause hypersensitivity to substrate analogs (38). This pocket appears to hold the incoming deoxynucleotide in close proximity to the active site, for coordination with one of the Mg^{2+} ions. In addition, Motif 2 residues K626 and R631 (from RT fingers), together with K902 from motif D (RT palm), may interact with both the sugar ring and phosphate groups, and provide stacking interactions with bases of the incoming deoxynucleotide. These interactions likely stabilize the telomeric DNA substrate during catalysis (21). The relatively conserved residues C931 and G932 from the RT palm define a "primer grip" (motif E) (29), which is essential for proper maintenance of telomeric DNA within telomerase active site (Fig. S3*B*). Additionally, R972 and K973 from the RT thumb, both located on an α-helix that packs into the minor groove of the RNA:DNA heteroduplex, zinteract with the DNA backbone (Fig. S3*C*). These residues, present also in telomerases that lack the TEN domain, may contribute significantly to repetitive addition processivity.

The interactions between telomerase and telomeric DNA are mediated by a variety of critical residues grouped into motifs characteristic for this family of polymerases (1, 12, 13). The spatial arrangement of these motifs resembles the shape of a double-stranded nucleic acid helix (21). The recently described "motif 3" of the reverse transcriptase domain provides several residues that may interact directly with telomeric DNA (39). Notably, mutations of motif 3 residues (V658A and K659A from RT fingers, and R669A from RT palm) cause telomerase hyperactivity (39), apparently resulting from weaker interaction with the double-stranded heteroduplex, facilitating telomeric DNA release after reaction. In our model, the positively charged K659 and R669 side chains are directed toward both DNA and RNA backbones, compatible with their essential contribution to nucleic acid binding (Fig. S3*A*). Increased repeat addition rate reduces processivity, however, possibly because the telomeric DNA cannot be stabilized sufficiently while the template-carrying RNA is realigned and prepared for the next reaction cycle (39). In support of this, Xie et al. (39) were able to obtain hyperactive and hyperprocessive human telomerase mutants by combining the V658A mutation with the deletion of residues 643–649 in the RT fingers and the hTR-U57C substitution in the RNA. The loop containing amino acid residues 643–649 precedes the motif 3 α-helix and is an intriguing structural feature: It may freely interact with both the telomeric DNA and RNA and likely stabilizes the position of the DNA substrate in the telomerase complex (Fig. S3*D*). Deletion of the 643–649 loop weakens this interaction and likely allows for more rapid dissociation of the heteroduplex, making the template binding site available for the next substrate deoxynucleotide and the next round of synthesis. Additionally, the hTR-U57C substitution results in extension of the RNA:DNA heteroduplex by an additional base, which could potentially

form a classical Watson–Crick pair, possibly further stabilizing the telomerase–telomeric DNA association.

In contrast to *T. castaneum* telomerase (PDB ID code 3KYL), the human protein contains two additional α-helices (residues 415–456) within the TRBD domain (similar to *T. thermophila*; PDB ID code 2R4G, residues 333–371), which, according to our model, together with the final α-helix in TRBD, form a three-helix bundle that packs tightly against the RT fingers (Fig. S4). This structural feature additionally stabilizes the interaction of RT and TRBD.

Essential N-Terminal Domain. The TEN domain, composed of a central β-sheet flanked by α-helices on both sides, is the most divergent domain within the telomerases (1, 12, 13). Nevertheless, it appears to be essential for proper telomere maintenance, because it "anchors" telomeric DNA (25). A recent study used a combination of comparative modeling and machine learning to identify several residues in TEN that are likely to play a role in nucleic acid binding (40). The sequence diversity among TEN domains of different species may be related to differences in telomeric DNA repeat sequences. TEN recognizes telomeric DNA in a sequence-specific manner, and several experiments have revealed differences in DNA binding affinity to different telomeric repeat sequences (41, 42). The TEN domain is separated from TRBD by a linker region, predicted to be largely unstructured, ranging in length from 20 to more than 500 amino acids, depending on the species (1). The TEN domain is believed to contribute to the processivity of the enzyme, because several studies have identified mutations or deletions of TEN residues that lead to a reduced ability of the telomerase to synthesize more than one telomeric repeat (23, 43). The crystal structure of the TEN domain from *T. thermophila* (PDB ID code 2B2A) (31) revealed certain features that adapt it for interaction with telomeric DNA. In particular, a deep cleft in the TEN domain surface is closely complementary to the shape of a double helix. Mutation of Q169, which is located in the central part of the cleft, compromises human telomerase processivity by hindering proper protein-DNA interaction (42, 44). In our modeled human TEN, Q169 forms a hydrogen bond with the backbone carbonyl group of P174 and the backbone amine group of L175, stabilizing the intervening loop, which establishes hydrogen bonds with another element of the TEN structure (Fig. S5). These interactions thus bridge adjacent structural elements and stabilize the overall shape of this region. The cleft is flanked by α-helical extensions on both sides and engages the RNA:DNA double helix, fitting into its major groove (Fig. 1). Jurczyluk et. al. (45) recently performed mutational analyses on TEN. Two mutations of particular interest involve residues 8–13 and 170–175 of TEN. The former exhibits wild type processivity but a decreased K_m, and the latter has significantly decreased processivity and an increased K_m. In our model, residues 8–13 make close contact with the heteroduplex, supporting this observation of decreased affinity. Residues 170–175 do have heteroduplex interactions in the model but are also partly buried in the TEN domain.

Interaction with the RNA Template. According to our model the template-carrying RNA component (TER) interacts with the TRBD, RT, and TEN domains of telomerase. Human TER is a structurally complex RNA molecule of 451 nucleotides, containing several conserved sequence and structural motifs (46). A characteristic pseudoknot domain, located in close proximity to the template repeat sequence and to an RNA loop domain designated CR4-CR5, is essential for telomerase catalytic activity (47). A 3'-terminal "H/ACA box" in TER contributes to the assembly and maturation of the ribonucleoprotein complex (48). The RNA pseudoknot region was shown to interact with a C-terminal region (residues 150–159) of TEN, providing further insight into the localization of the template region within the human telomerase complex (49). Our modeled structure of TRBD exposes a wide cleft opening toward the C terminus of TEN, which might bind the pseudoknot domain. Furthermore, the surface of the cleft presents several lysines (K492, K493, K511) that could interact with RNA. The TER CR4-CR5 domain has been shown to bind to the CP and T motifs of the TRBD domain (1), which are located at a considerable distance (approximately 23 Å) from the putative pseudoknot-binding site in our model. Recent work by Egan and Collins provides insight into hTERT-hTR interactions and will be useful for future studies that attempt to model the full length hTR (50).

RT:TRBD-TEN Interaction. The shape of the TEN domain restricts its possible orientation with respect to the other domains and the RNA:DNA heteroduplex. Mutual positioning of the telomeric DNA substrate and RT, to ensure proper interactions within the reverse transcriptase active site, determines the distance between the surface of TRBD and the major groove of the bound double helix. Therefore, the possible TEN orientations are dramatically limited. Together, these constraints aided in the assembly of our model of the human telomerase RNP complex. The surfaces of RT, TRBD, and TEN expose poorly conserved residues, hindering the modeling of interactions between TEN and the other two domains. However, detailed analysis of surface residues in the model reveals increased conservation of uncharged amino acids at the domain interfaces (e.g., G100, F101 in TEN and G967, V1025 in the RT thumb; T117, S118 in TEN and S504, L505, A542 in TRBD; see Fig. S6). Despite poor surface conservation in TEN, the proposed interface represents an optimal structural fit between TEN and RT:TRBD and the RNA:DNA heteroduplex. Additional support for our proposed assembly is provided by Sealey et al. (42), who reported that T116A, T117A, and S118A mutations in TEN compromise repetitive addition processivity but do not alter DNA binding affinity and thus probably affect interaction with the remaining protein domains of the telomerase. In the model, these three residues are located at the binding surface between TEN and the TRBD and RT thumb domains. This interface is far from the nucleotide binding region, with the closest nucleotide atom approximately 16 Å away. It is, however, involved in the slower motions, presumed to be functionally relevant (see below), and this may explain the role of these residues in the processivity.

Mechanism of Telomerase Action. To investigate the processive mechanism of telomerase, we used elastic network models to generate the mechanistic step shown in close-up in Fig. 2. We utilized our Anisotropic Network Model (ANM) (51), with one node for each C^{α}, P and O4′ atom, and with identical springs placed between pairs of these atoms within 13 Å. To generate the processive conformations shown, we used the global mode of motion to move the original structure. This reveals in detail how the telomerase structure effects the motion of the template. Views of the whole structure are shown in further detail in Fig. S7 and in Movies S1 and S2.

Our 3D model assumes that the intrinsic template-carrying RNA molecule forms a heteroduplex with the single-stranded telomeric DNA, while the architecture of both the RT active site and the TEN anchor domain are adapted for double-stranded nucleic acid binding. The template-encoding region of TER contains one complete 6-nt repeat complementary to the human telomeric $(TTAGGG)_n$ repeat sequence, which can initiate RNA:DNA heteroduplex formation. Multiple sequence alignment of telomerase RNAs from several phyla (extracted from the Telomerase Database, http://telomerase.asu.edu) (52) reveals that the telomere repeat template sequence is partially repeated, extending the potential length of the helical heteroduplex region, as was previously proposed (53). In the human enzyme, such an extended RNA:DNA heteroduplex would contain at least 10

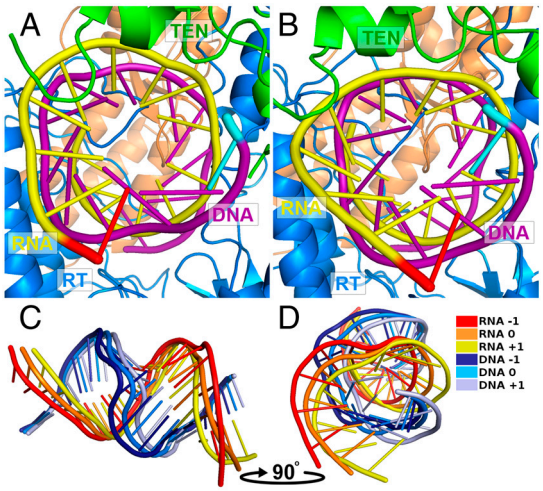


Fig. 2. Structural model for the processive motion of human telomerase. The procession of the DNA:RNA heteroduplex is a critical aspect of telomerase function. Rotation and translation of the heteroduplex is evident in the global mode of the elastic network. The effects of following the global mode in the (*A*) negative (−1) and (*B*) positive (+1) directions are shown. Termini closest to the viewer are highlighted: the 3′ end of RNA in red and the 5′ end of DNA in cyan. See Fig. S7 for corresponding views of the entire structure and Movies S1 and S2 for two dynamic views of this motion. *C* offers a side view and *D* a face view of the heteroduplex for three states of the negative global mode (−1), the original state (0), and the positive mode deformation (+1).

base pairs, exactly filling the space between the RT active site and the experimentally supported nucleic acid binding region of TEN (Fig. 1). Thus, our model provides a strong structural basis for the mechanism shown in Fig. 2, in which, after completion of a single telomere repeat synthesis cycle, the template RNA is moved ahead at the polymerase site to the next base. The TER template region must then shift relative to the substrate DNA, while maintaining the RNA:DNA heteroduplex (Fig. S8). The second, partial repeat (5′-UAAC-3′) adjacent to the TER template region complements the newly synthesized telomeric DNA repeat, while the TER template region itself (5′-UAACCC-3′) forms a 5′ overhang ready for the next reaction cycle. Our model strongly supports an essential role of TEN in stabilizing the assembling heteroduplex in an orientation that promotes proper interaction with the RT active site. Further, our dynamics simulations suggest the important role of the structure in controlling the shift along the template to the next base to be copied. The current model does not provide insight into the mechanism by which the nucleic acid translocates after the complete synthesis of each complete template sequence; however, there is the possibility that the protein can extend together with the RNA through the repeated synthesis steps for one template cycle before recoiling to activate and reposition the template for the next cycle.

Notably, the template-encoding region of TER is followed by a sequence rich in uracil (U) residues that are capable of forming wobble base pairs with guanine (G). Therefore, the proposed RNA:DNA heteroduplex formed during reverse transcription may actually be longer, to help stabilize the helical structure during the RNA template realignment step between reaction cycles. Such a possible extended helix could contribute significantly to the extension mechanism.

Several telomerases (e.g., those from *C. elegans* and *T. castaneum*) lack a distinguishable TEN domain, which is essential for activity and processivity of the human and *Tetrahymena* enzymes (1, 12, 13). Despite this, the *T. castaneum* enzyme appears to be active in vitro (29) and genetic evidence suggests that the *C. elegans* telomerase is capable of synthesizing multiple telomeric repeats in vivo (54). Meier and colleagues found that TRBD domain in *C. elegans* is preceded by a domain that might be distantly homologous to TEN (54); however, our methods failed to detect significant similarity of this region to any known protein domains. The RT thumb and IFD motif (RT fingers) have been shown to play roles in repetitive addition processivity (20, 25) and, in the absence of TEN, could provide substrate stabilization during subsequent reaction cycles. Species-specific accessory factors that influence processivity could also compensate for the lack of a TEN domain (12, 13). Interestingly, we found that telomerases lacking the TEN domain (e.g., PDB ID code 3KYL) also lack the α-helical insertion within the TRBD (Fig. S4), which would allow for a more elastic RT:TRBD interface (discussed above). Telomerases possessing TEN domains possibly do not require such elasticity because TEN aids the nucleic acid binding and regulates the processivity.

Conclusions

The availability of a structural model of the assembled human telomerase complex presented here provides information necessary for investigating its mechanism further, as well as for locating its interactions within the complex cellular signaling networks in which it is known to participate (8, 17, 18). A complete structural model of telomerase may also accelerate the development of new anticancer therapies that aim to abolish telomerase activity in proliferating tumor cells, or to augment enzymatic activity in cases of telomerase insufficiency diseases.

Methods

We present a brief overview of our methods here with further details given in *SI Methods*.

We have utilized the human telomerase protein sequence (GenBank accession no. NM_198253.2) and PSI-BLAST (55) to study the sequence conservation within the telomerase protein family. Multiple sequence alignments of collected sequences were prepared with PCMA (56), while PSI-PRED (57) was used for secondary structure prediction. Templates for comparative modeling of human telomerase domains were identified from the full-length sequence and individual domains using the Gene Relational Database (GRDB) system, which stores pre-calculated Meta-BASIC mappings (32) between Pfam families, conserved domains, and PDB structures. The results were validated by 3D-Jury (58) and manual inspection followed by 3D assessment (59).

Three-dimensional models of human telomerase protein domains were generated with Modeller (60) based on manually curated, high confidence sequence-to-structure alignments. These models were built separately for (*i*) the TEN domain, using PDB ID code 2B2A (31); (*ii*) the RT:TRBD subcomplex, using PDB ID code 3KYL (29) and the superimposed TRBD domain from PDB ID code 2R4G (30) as templates. The resulting 3D models were then assembled manually after careful consideration of the CABS (61) results for protein domain docking, published experimental data (Table S1), and conservation of surface residues from ConSurf (62) (Fig. S6). Assembly of the modeled domains proceeded by first rigidly docking using an exhaustive global search in a six-dimensional space of "ligand" rotations and translations against the frozen structure of the "receptor" using FTDOCK (63). The resulting 10,000 FTDOCK top-scoring structures were rescored with the CABS force field and grouped using hierarchical clustering. From each cluster, a representative with the lowest energy was selected, leaving 30 models.

Positions of the intrinsic RNA template and single-stranded telomeric DNA substrate in the human telomerase model were copied from the *T. castaneum* telomerase structure (PDB ID code 3KYL) after superposition of their RT and TRBD domains. The 3D partial model comprising all three protein domains and the RNA:DNA heteroduplex was energy-minimized with Tripos SYBYL using an AMBER force field (64), followed by a short molecular dynamics run to relieve steric clashes and improve internal packing.

The final model was used to investigate the motions of the structure using coarse-grained elastic network models in conjunction with normal mode analysis. This approach has been widely used to investigate important functional motions of biomolecular structures (65). Notably the computed motions are quite insensitive to the details of the structure, which means the computed motions reported here are robust and unlikely to be changed by any minor errors in the model.

ACKNOWLEDGMENTS. We thank the reviewers for their insightful suggestions and Peter Zaback for critical reading of the manuscript. This work was supported by EMBO Installation, National Science Foundation Grant MCB-1021785, National Institutes of Health grants R01GM072014 and R01GM081680, Foundation for Polish Science (Focus, Team), Ministry of Science and Higher Education (N N301 159435) and European Social Fund (UDA-POKL.04.01.01-00-072/09-00), and Center for Integrated Animal Genomics (Iowa State University) grants.

1. Autexier C, Lue NF (2006) The structure and function of telomerase reverse transcriptase. *Annu Rev Biochem* 75:493–517.
2. Blackburn EH, Collins K (2010) Telomerase: An RNP enzyme synthesizes DNA. *Cold Spring Harb Perspect Biol*, 10.1101/csshperspect.a003558.
3. McEachern MJ, Krauskopf A, Blackburn EH (2000) Telomeres and their control. *Annu Rev Genet* 34:331–358.
4. Lansdorp PM (2009) Telomeres and disease. *EMBO J* 28:2532–2540.
5. Armanios M (2009) Syndromes of telomere shortening. *Annu Rev Genomics Hum Genet* 10:45–61.
6. Murnane JP (2006) Telomeres and chromosome instability. *DNA Repair (Amst)* 5:1082–1092.
7. Flores I, Blasco MA (2010) The role of telomeres and telomerase in stem cell aging. *FEBS Lett* 584:3826–3830.
8. Aubert G, Lansdorp PM (2008) Telomeres and aging. *Physiol Rev* 88:557–579.
9. Artandi SE, DePinho RA (2010) Telomeres and telomerase in cancer. *Carcinogenesis* 31:9–18.
10. Calado RT, Young NS (2009) Telomere diseases. *N Engl J Med* 361:2353–2365.
11. Mitchell JR, Wood E, Collins K (1999) A telomerase component is defective in the human disease dyskeratosis congenita. *Nature* 402:551–555.
12. Wyatt HD, West SC, Beattie TL (2010) InTERTpreting telomerase structure and function. *Nucleic Acids Res* 38:5609–5622.
13. Sekaran VG, Soares J, Jarstfer MB (2010) Structures of telomerase subunits provide functional insights. *Biochim Biophys Acta* 1804:1190–1201.
14. Gonzalez-Suarez E, et al. (2001) Increased epidermal tumors and increased skin wound healing in transgenic mice overexpressing the catalytic subunit of telomerase, mTERT, in basal keratinocytes. *EMBO J* 20:2619–2630.
15. Dudognon C, et al. (2004) Death receptor signaling regulatory function for telomerase: hTERT abolishes TRAIL-induced apoptosis, independently of telomere maintenance. *Oncogene* 23:7469–7474.
16. Santos JH, Meyer JN, Van Houten B (2006) Mitochondrial localization of telomerase as a determinant for hydrogen peroxide-induced mitochondrial DNA damage and apoptosis. *Hum Mol Genet* 15:1757–1768.
17. Epel ES, et al. (2010) Dynamics of telomerase activity in response to acute psychological stress. *Brain Behav Immun* 24:531–539.
18. Wolkowitz OM, et al. (2011) Resting leukocyte telomerase activity is elevated in major depression and predicts treatment response. *Mol Psychiatry*, 10.1038/mp.2010.133.
19. Greider CW, et al. (1985) Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell* 43:405–413.
20. Hossain S, Singh S, Lue NF (2002) Functional analysis of the C-terminal extension of telomerase reverse transcriptase. A putative "thumb" domain. *J Biol Chem* 277:36174–36180.
21. Gillis AJ, Schuller AP, Skordalakes E (2008) Structure of the Tribolium castaneum telomerase catalytic subunit TERT. *Nature* 455:633–637.
22. Chen JL, Greider CW (2003) Determinants in mammalian telomerase RNA that mediate enzyme processivity and cross-species incompatibility. *EMBO J* 22:304–314.
23. Lue NF (2004) Adding to the ends: What makes telomerase processive and how important is it? *Bioessays* 26:955–962.
24. Latrick CM, Cech TR (2010) POT1-TPP1 enhances telomerase processivity by slowing primer dissociation and aiding translocation. *EMBO J* 29:924–933.
25. Lue NF (2005) A physical and functional constituent of telomerase anchor site. *J Biol Chem* 280:26586–26591.
26. Lingner J, Price C (2009) Conservation of telomere protein complexes: Shuffling through evolution. *Crit Rev Biochem Mol Biol* 44:434–446.
27. Cohen SB, et al. (2007) Protein composition of catalytically active human telomerase from immortal cells. *Science* 315:1850–1853.
28. Fu D, Collins K (2007) Purification of human telomerase complexes identifies factors involved in telomerase biogenesis and telomere length regulation. *Mol Cell* 28:773–785.
29. Mitchell M, Gillis A, Futahashi M, Fujiwara H, Skordalakes E (2010) Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat Struct Mol Biol* 17:513–518.
30. Rouda S, Skordalakes E (2007) Structure of the RNA-binding domain of telomerase: Implications for RNA recognition and binding. *Structure* 15:1403–1412.
31. Jacobs SA, Podell ER, Cech TR (2006) Crystal structure of the essential N-terminal domain of telomerase reverse transcriptase. *Nat Struct Mol Biol* 13:218–225.
32. Ginalski K, von Grotthuss M, Grishin NV, Rychlewski L (2004) Detecting distant homology with Meta-BASIC. *Nucleic Acids Res* 32:W576–581.
33. Sarafianos SG, et al. (2001) Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. *EMBO J* 20:1449–1461.
34. Lingner J, et al. (1997) Reverse transcripase motifs in the catalytic subunit of telomerase. *Science* 276:561–567.
35. Weinrich SL, et al. (1997) Reconstitution of human telomerase with the template RNA component hTR and the catalytic protein subunit hTRT. *Nat Genet* 17:498–502.
36. Bosoy D, Peng Y, Mian IS, Lue NF (2003) Conserved N-terminal motifs of telomerase reverse transcriptase required for ribonucleoprotein assembly in vivo. *J Biol Chem* 278:3882–3890.
37. Drosopoulos WC, Prasad VR (2007) The active site residue Valine 867 in human telomerase reverse transcriptase influences nucleotide incorporation and fidelity. *Nucleic Acids Res* 35:1155–1168.
38. Smith RA, Anderson DJ, Preston BD (2006) Hypersusceptibility to substrate analogs conferred by mutations in human immunodeficiency virus type 1 reverse transcriptase. *J Virol* 80:7169–7178.
39. Xie M, Podlevsky JD, Qi X, Bley CJ, Chen JJ (2009) A novel motif in telomerase reverse transcriptase regulates telomere repeat addition rate and processivity. *Nucleic Acids Res* 38:1982–1996.
40. Lee JH, et al. (2008) Striking similarities in diverse telomerase proteins revealed by combining structure prediction and machine learning approaches. *Pac Symp Biocomput* 13:501–512.
41. Wyatt HD, Lobb DA, Beattie TL (2007) Characterization of physical and functional anchor site interactions in human telomerase. *Mol Cell Biol* 27:3226–3240.
42. Sealey DC, et al. (2010) The N-terminus of hTERT contains a DNA-binding domain and is required for telomerase activity and cellular immortalization. *Nucleic Acids Res* 38:2019–2035.
43. Zaug AJ, Podell ER, Cech TR (2008) Mutation in TERT separates processivity from anchor-site function. *Nat Struct Mol Biol* 15:870–872.
44. Wyatt HD, Tsang AR, Lobb DA, Beattie TL (2009) Human telomerase reverse transcriptase (hTERT) Q169 is essential for telomerase function in vitro and in vivo. *PLoS One* 4:e7176.
45. Jurczyluk J, et al. (2011) Direct involvement of the TEN domain at the active site of human telomerase. *Nucleic Acids Res* 39:1774–1788.
46. Feng J, et al. (1995) The RNA component of human telomerase. *Science* 269:1236–1241.
47. Theimer CA, Feigon J (2006) Structure and function of telomerase RNA. *Curr Opin Struct Biol* 16:307–318.
48. Collins K (2008) Physiological assembly and activity of human telomerase complexes. *Mech Ageing Dev* 129:91–98.
49. Moriarty TJ, Marie-Egyptienne DT, Autexier C (2005) Regulation of 5′ template usage and incorporation of noncognate nucleotides by human telomerase. *RNA* 11:1448–1460.
50. Egan ED, Collins K (2010) Specificity and stoichiometry of subunit interactions in the human telomerase holoenzyme assembled *in vivo*. *Mol Cell Biol* 30:2775–2786.
51. Atilgan AR, et al. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* 80:505–515.
52. Podlevsky JD, Bley CJ, Omana RV, Qi X, Chen JJ (2008) The telomerase database. *Nucleic Acids Res* 36:D339–343.
53. Gilley D, Lee MS, Blackburn EH (1995) Altering specific telomerase RNA template residues affects active site function. *Genes Dev* 9:2214–2226.
54. Meier B, et al. (2006) trt-1 is the *Caenorhabditis elegans* catalytic subunit of telomerase. *PLoS Genet* 2:e18.
55. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
56. Pei J, Sadreyev R, Grishin NV (2003) PCMA: Fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics* 19:427–428.
57. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292:195–202.
58. Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: A simple approach to improve protein structure predictions. *Bioinformatics* 19:1015–1018.
59. Ginalski K, Rychlewski L (2003) Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins* 53(Suppl 6):410–417.
60. Eswar N, et al. (2006) Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* (John Wiley & Sons, Inc., NY), Chapter 5:Unit 5.6.
61. Kolinski A (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol* 51:349–371.
62. Landau M, et al. (2005) ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* 33:W299–302.
63. Jackson RM, Gabb HA, Sternberg MJ (1998) Rapid refinement of protein interfaces incorporating solvation: Application to the docking problem. *J Mol Biol* 276:265–285.
64. Cornell WD, et al. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 117:5179–5197.
65. Bahar I, Lezon TR, Yang LW, Eyal E (2010) Global dynamics of proteins: Bridging between structure and function. *Annu Rev Biophys* 39:23–42.